

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Ottawa, Canada, 16-18 May 2005)

Topic (v): Quality indicators and quality reporting

**THE CONTRIBUTION OF DIFFERENT WAYS OF DEALING WITH NON-RESPONSES IN
FRENCH BUSINESS SURVEYS**

Invited Paper

Submitted by INSEE, France¹

Abstract: French business surveys are generally done by mail. The return of the questionnaires is spread out over a long period. In the end, the non-responses are treated in different ways according to the size of the units. For small enterprises, reminder letters are sent, and the final non-responses are treated using classic statistical methods, such as reweighting or imputation. For large enterprises, statisticians try to get the information through the visit of an enumerator, and then through the use of fiscal data, which are less complete than the data of the statistical questionnaire.

This paper gives quality elements about these stages of the treatment of business surveys, from the point of view of the producer: what is the impact of these stages on the precision, is there an optimal strategy, from a budget point of view, concerning the choices relative to the treatment of non-responses?

1. INTRODUCTION

1. Conducting business surveys is a long and difficult task, implying various stages of development. In France, these surveys are generally done by mail, with the return of the questionnaires spread out over a long period. This paper will focus on the main business surveys, the annual enterprise survey [4], and study the contribution of some stages of the treatment of the non-responses.

2. Different kinds of treatment are used, depending on the size of the non-respondents: for small enterprises, reminder letters are sent to the non-respondents to request their questionnaires, while large enterprises are visited by specialized enumerators, which is much more costly.

3. The sampling plan of the annual enterprise surveys includes an exhaustive part (large enterprises) and sampled strata (figure 1). To simplify, the stratification uses two criteria: the economic sector to which the enterprise belongs, and its size (defined using the number of employees).

4. Another feature of the survey is that many of the variables of the questionnaire are accounting variables: these variables are also available in some fiscal files² (annual income statements of the enterprises), which the fiscal administration agrees to transmit to INSEE. The recent improvement to the

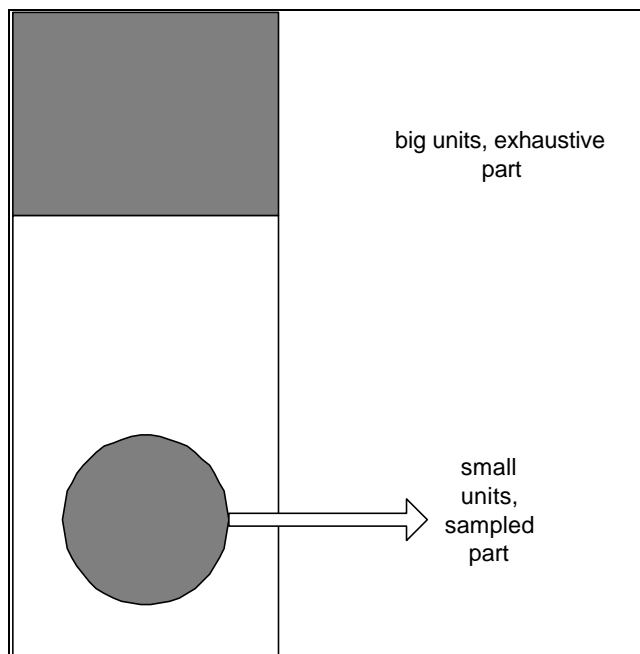
¹ Prepared by Philippe Brion (philippe.brion@insee.fr).

² According to the French « Plan comptable général », which asks enterprises to use the same concepts for different uses, particularly fiscal and statistical uses.

date of availability of these files led statisticians to use these data in a more important way, especially for large enterprises: however, the fiscal data, as it will be presented hereafter, do not give all the variables of the statistical questionnaire, and cannot replace it in a direct way.

5. So, for the large non-respondents, we try to get the data by the visit of an enumerator, and, if this is not possible, we use the fiscal data for the accounting data, and for other data, we have to impute a value. One important variable is the code of economic activity (in French APE : *Activité Principale de l'Entreprise*), resulting in the breakdown of the turnover in elementary activities. The information concerning this breakdown is available only in the statistical questionnaire, and allows classifying the enterprise in a correct way (according to the NACE nomenclature), which is essential to produce statistics about the shares of turnover of the different economic sectors, for example. If only the fiscal data are available, we have to use an approximate value of the APE code, which is the value that was used in the sampling frame (built from the business register). Then, if a large enterprise is classified in a "wrong" class (because of the "old" value), the impact on some estimations may be important.

Figure 1 : Sampling plan of the annual enterprise surveys



6. For small enterprises, dealing with non-responses is done by more classic statistical methods: imputation using the data of the former year, if the enterprise was already in the sample, and an evolution rate, or selective hot-deck (within defined categories). To simplify, we will use in the following sections a theoretical framework where a reweighting method is used, to evaluate the impact of this stage.

7. So the precision, in the end, depends on different elements: sampling plan, treatment of large and small non-responses, and data editing. Many papers have been written about the question of the efficiency of the data editing process (see for example [2]), and on selective editing methods (see for example [3]). Here, we will not focus on this question (for details on the methods used for the French annual enterprise surveys, see [4]), and we will consider the data obtained with the processing chain as "perfect". We will focus on the problems resulting from non-responses: what kinds of actions have to be considered as priorities for the statistician? Then, it is important to quantify the impact on the precision of different stages (sampling plan, different kinds of treatment of the non-responses), in order to be able to define a strategy relative to the non-responses.

8. The first part of this paper presents the simplified model that is used. The second part deals with the “reminder choices”, which are shared between reminder letters for small enterprises, and visits of large enterprises (more costly). So, this question concerns the cost aspects, but may also be considered from a timeliness point of view: how to produce results with a shorter delivery time? This paper will focus on the estimate of aggregates (for example turnover) for different domains (for example activity sectors defined as the level “three digits” of the NACE).

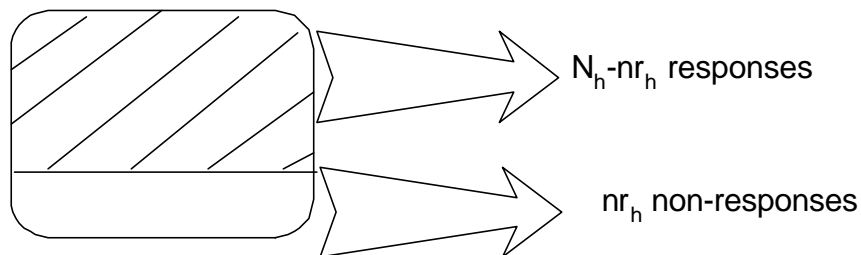
II. THEORETICAL BACKGROUND

9. The estimation of an aggregate (for example turnover) for an economic sector will be obtained as the sum of the estimates of turnover on all strata, multiplied by the variable having the value « 1 » if the enterprise belongs to the sector, « 0 » if not, which will be written 1_{APE} (APE is the code of economic activity). One cannot only just add the values obtained on the strata « belonging » to this economic sector, since for many enterprises, the main activity may be different from the value used in the sampling frame: having a correct classification of enterprises is one of the main results of the survey. In the following parts, notation h will be used for the strata of large enterprises, and k for those of small enterprises.

A. Estimator used for large enterprises

10. For each stratum h , questionnaires are sent to all N_h enterprises.

Figure 2: stratum of large enterprises



We get $N_h - nr_h$ responses, if the number of non-responses is nr_h .

The contribution of this stratum to the estimator of the total of a variable Y (inside an economic sector) is :

$$\sum_{N_h - nr_h} Y_i * 1_{APE_{surv}}(i) + \sum_{nr_h} Y_i * 1_{APE_{reg}}(i)$$

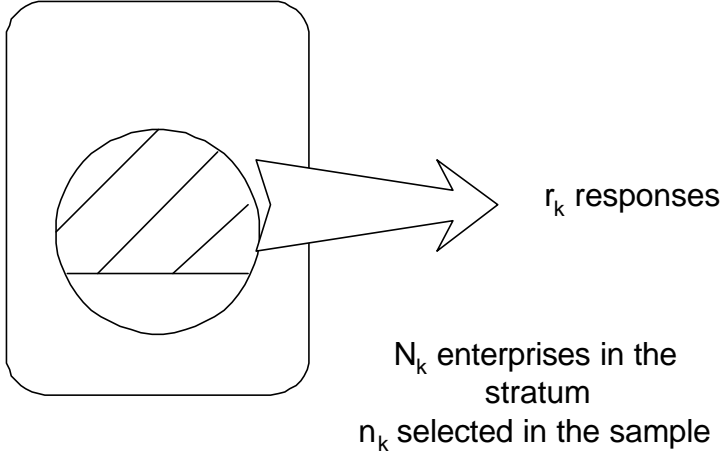
where $1_{APE_{surv}}$ is the variable of value 1 if the enterprise belongs to the economic sector according to the result of the survey, 0 else,

and $1_{APE_{reg}}$ is the variable of value 1 if the enterprise belongs to the sector according to the value of the sampling frame (business register).

So, for the non-responses, an “old” value has to be used for classifying the units in economic sectors, and this has consequences on the quality of estimates.

B. Estimator used for small enterprises

Figure 3: sampled stratum



11. In each stratum k , n_k enterprises were selected in the sampled, and r_k responses were obtained.

12. Under the hypothesis that the non-response behaviour is homogeneous within each stratum, for the contribution of the stratum to the estimation of the total of a variable Y , it is possible to use a re-weighted estimator:

$$\frac{N_k}{r_k} \sum_{rk} Y_i * 1_{APEsurv}(i).$$

C. Bias and variance of the estimator of the total

The bias is due only to large enterprises

13. Under the hypothesis that the "choice" of the nr_h units non-respondents, among large enterprises, is random within the stratum h , the error relative to the estimation of the total and due to this stratum is:

$$\sum_{nrh} Y_i * 1_{APEreg}(i) - \sum_{nrh} Y_i * 1_{APEsurv}(i)$$

The expected value of this error, that is the bias, is:

$$\frac{nr_h}{N_h} \sum_{N_h} Y_i * (1_{APEreg} - 1_{APEsurv})(i)$$

and may be estimated using the « sample » of respondents by:

$$\frac{nr_h}{N_h - nr_h} \sum_{N_h - nr_h} Y_i * (1_{APEreg} - 1_{APEsurv})(i)$$

On all strata, the bias may then be expressed as a function of the non-responses among the big units $\sum_h nr_h^* A_h$.

The two categories (small and large enterprises) have a contribution to the variance

Small enterprises

14. For each stratum k , the variance of the estimator of the total is: $\frac{N_k^2}{r_k} (1 - \frac{r_k}{N_k}) S_k^2$, again under the hypothesis that the non-response behaviour is homogeneous within the stratum, where S_k^2 is the « corrected » variance of the variable $Y_i^* 1_{APE_{Surv}}(i)$ on stratum k .

Large enterprises

15. The estimator of the total for the stratum h may be written as:

$$\sum_{N_h} Y_i^* 1_{APE_{Surv}}(i) + \sum_{nr_h} Y_i^* (1_{APE_{Reg}} - 1_{APE_{Surv}})(i)$$

The first term being constant, the variance of this estimator of the total for stratum h is:

$$V(\sum_{nr_h} Y_i^* (1_{APE_{Reg}} - 1_{APE_{Surv}})(i)) = nr_h (1 - \frac{nr_h}{N_h}) B_h^2$$

where :

B_h^2 is the « corrected » variance of the variable $Y_i^* (1_{APE_{Reg}} - 1_{APE_{Surv}})(i)$

and may be estimated with the $N_h - nr_h$ « responses ».

D. Total mean square error

16. Adding the square of the bias and the variance that was obtained on all strata (small and large enterprises) gives the value of the mean square error:

$$(\sum_h nr_h A_h)^2 + \sum_h (nr_h - \frac{nr_h^2}{N_h}) B_h^2 + \sum_k \frac{N_k^2}{r_k} (1 - \frac{r_k}{N_k}) S_k^2$$

the first and the second term being added on the H strata of large enterprises, and the third one on the K strata of small enterprises.

III. CONSEQUENCES ON THE REMINDER PROCESS

A. Minimization of the mean square error under some constraints

17. The statistician may have to act on two series of parameters: nr_h , the number of non-responses in each stratum of big units (with visits of enumerators), and r_k , the number of respondents in strata of small units (with follow-up letters or telephone calls).

18. The visits of enumerators have a cost C_1 and the follow-up letters (or the telephone calls) a cost C_2 (cheaper).

19. The general budget constraint is as follows:

$$\sum_k C_1(nr_{h0} - nr_h) + \sum_h C_2(r_k - r_{k0}) = C$$

where nr_{h0} is the number of big non-responses that was obtained « initially », and r_{k0} is the number of respondents, among small units, obtained « initially ».

20. The quantity to minimize is the mean square error:

$$\left(\sum_h nr_h A_h\right)^2 + \sum_h \left(nr_h - \frac{nr_h^2}{N_h}\right) B_h^2 + \sum_k \frac{N_k^2}{r_k} \left(1 - \frac{r_k}{N_k}\right) S_k^2$$

with $0 \leq nr_h \leq nr_{h0}$

and $r_{k0} \leq r_k \leq n_k$.

B. Solution

21. By means of the Lagrange multiplier:

$$L = \left(\sum_h nr_h A_h\right)^2 + \sum_h \left(nr_h - \frac{nr_h^2}{N_h}\right) B_h^2 + \sum_k \frac{N_k^2}{r_k} \left(1 - \frac{r_k}{N_k}\right) S_k^2 + I \left(\sum_h C_1 nr_h - \sum_k C_2 r_k\right)$$

$$\frac{\partial L}{\partial r_k} = \frac{S_k^2}{r_k^2} N_k^2 + I C_2 = 0$$

$$\frac{\partial L}{\partial nr_h} = 2\left(\sum_h nr_h A_h\right) A_h + B_h^2 \left(1 - 2 \frac{nr_h}{N_h}\right) + I C_1 = 0$$

22. The complete solution is presented in the annex: the budget constraint leads to a first sharing out between the total budget for the visits and that for the follow-up letters; then, within each category (large/small enterprises), we obtain the number of visits or follow-up letters for each stratum (for the follow-up letters, it will obviously be a Neyman allocation relative to the “budget given” to the strata of small units).

23. From a practical point of view, how do we use the previous considerations? This is done by using quantified values of the different elements A_h , B_h , S_k , obtained with the data from the former survey. The more or less important stability (relative to classifying the enterprises into economic sectors) of each stratum of large enterprises will thus be taken into account, as well as the homogeneity of each stratum of small enterprises relatively to the studied variable.

IV. CONCLUSION

24. The conceptual framework that has been presented provides elements for the implementation of a strategy of separate treatments of the non-responses, even if some strong hypotheses were introduced. It can be redefined by anticipating a « success rate » in the different kinds of reminder actions (this may be introduced in the elementary costs). Other further developments could be relative to variables that are not accounting variables (and so not available in the fiscal data): for them, the value is imputed, for a big unit « non-respondent », using a ratio computed on the respondents and applied to a fiscal variable.

25. The approach developed in this paper gives elements for the management of the non-responses, by quantifying on which strata it is important to increase the response rate, and it may be interesting to take into account the timeliness aspects. It does not aim at solving all problems relative to the production of results: other elements are important (see for example [1]), and the final checking of the results may also need other approaches, such as macro-editing, or the comparison with other sources.

References

- [1] Brion Ph. "The management of quality in French business surveys", European conference on quality and methodology in official statistics, Mainz, may 2004.
- [2] Granquist L., Kovar J., "Editing of survey data: how much is enough ?" , in *Survey measurement and process quality*, Lyberg et al. ed., J. Wiley and sons, New York, 1997.
- [3] Lawrence D., McKenzie R., "The general application of significance editing", *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253, 2000.
- [4] Rivière P., "The new annual enterprise surveys in France", *Courrier des statistiques*, English series n°3, 1997 (available on www.insee.fr at http://www.insee.fr/en/ffc/docs_ffc/cs78e.pdf).

ANNEX

MINIMIZATION OF THE VARIANCE OF THE ESTIMATOR UNDER CONSTRAINT

Subject to the fact that the constraints are not saturated, the values nr_h may be expressed as a function of I , taking the second set of equations of part 3.2. :

$$2(\sum_h nr_h A_h)A_h + B_h^2(1 - 2\frac{nr_h}{N_h}) + IC_1 = 0$$

By multiplying each member by $\frac{1}{2} \frac{A_h N_h}{B_h^2}$, and cumulating on all strata h :

$$(\sum_h nr_h A_h)(\sum_h \frac{N_h A_h^2}{B_h^2} - 1) = -\frac{IC_1}{2} \sum_h \frac{A_h N_h}{B_h^2} - \sum_h \frac{A_h N_h}{2}$$

$$\text{Since } \frac{S_k}{r_k} \frac{N_k^2}{2} + IC_2 = 0, \text{ it can be deduced } \frac{\sum_k S_k N_k}{\sum_k r_k} = \sqrt{-IC_2}$$

By using the two equations on I , it is possible to express $(\sum_h nr_h A_h)$ as a function of $\sum_k r_k$.

Using the equation at the beginning of this annex, nr_h may be expressed as a function of $(\sum_h nr_h A_h)$, and then, using the initial cost function, the quantified values of the different parameters may be computed.
