Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 341–347

Discussion

Philippe Brion[1]

I would like to congratulate Roderick Little on having produced an article which raises important questions for government statisticians with respect to the production of statistics: especially topics like small area estimation, estimation of the standard error of an estimate, or the study of nonsampling errors. These questions are indeed practical challenges statisticians are faced with, and on these subjects. Little makes a very interesting review of the answers presently put forth by official statisticians, characterized by the use of some classical methods. Little proposes a new class of approaches consisting of Bayesian calibrated methods to improve the quality of the results. Even if I feel that on some subjects, particularly the topic of the impact of imputation methods, some work has been done during the recent years, and that the view of R. Little may be rather pessimistic, I agree with his review of the methodological choices made at National Statistical Institutes (NSI). Several circumstances may explain this situation.

One first element of explanation is that government statisticians have to produce statistics that are considered impartial: since using models could lead to the results being considered influenced by *a priori* choices, official statisticians are, rightly or wrongly, more confident with "objective" methods. Many authors have raised this point, and I do not develop it further here.

Secondly, government statisticians are generally faced with multiple kinds of demands: demand for global aggregates as well as demands for very detailed information, these objectives sometimes being contradictory. For example, for business statistics, the demand may be for aggregates for the national accounts as well as for data on very specific populations, for example small enterprises of a given economic sector. The use of estimation procedures based on one set of weights, often obtained with model-assisted methods, gives consistency to the whole set of results through linear estimates, and at the same time gives some robustness to them (especially against model misspecification). The statistical products, often consisting of thousands of figures, or more, have to be published with an internal consistency, first of all with additive constraints (the results of regions, or of categories, adding for example to the total).

Thirdly, no production process within an NSI can be seen as independent, but rather as belonging to a whole device, where internal coherence has to be considered. For example, some large surveys may be conducted for structural purposes, and smaller ones are dedicated to specific topics. The statistical office has to manage an industrial process with some global coherence and rules to be applied by everyone, in contrast to using *ad hoc*

[1] INSEE, 18, Boulevard Adolphe Pinard, F-75675 Paris Cedex 14, France. Email: philippe.brion@insee.fr

methods for every process, even if those methods might be more efficient for each specific purpose.

Now, using the point of view of a production process manager, I elaborate on two subjects where important methodological questions are raised for government statisticians, especially since budget restrictions affect them. These two subjects are maybe at the very heart of the challenges discussed by Little in his article: the potential opened up by availability of multiple sources of data, and data editing of large databases. The first one is briefly discussed in Little's article (in Sec. 4.8), and regarding the second one it should be noted that Little has produced a paper (Little and Smith 1987) proposing an elegant method to handle the task.

1. The Use of Multiple Sources

As mentioned in Little's article, this question is one of the most challenging for government statisticians, especially since increased use of administrative data is considered a way to improve the quality of the statistics while at the same time decreasing the costs of production.

I will not discuss here all kinds of problems linked to this use (for example, concept differences between the administrative world and the statistical world), but rather focus on statistical estimation by utilizing a concrete example concerning business statistics.

INSEE, the French NSI, has during the last five years renewed its system of producing the structural business statistics by directly using tax data (annual statements of income sent by enterprises to the tax authorities) and "social data" (declarations sent to social security administration, giving information about salaries and wages). Since all the information needed to produce the structural business statistics is not available in the administrative sources, a statistical survey conducted on a sample of enterprises, completes the system of production, named ESANE (Brion 2011).

A part of the questionnaire of this statistical survey is dedicated to the different activities conducted by the enterprise. This information is used for the national accounts, but also to evaluate and if necessary to revise the value of the code of principal activity of the enterprise (in French, APE code, *Activité Principale de l'Entreprise*, referring to the nomenclature of activities). The value of this code is indeed available for all enterprises within the business register; however, it is revisited for the enterprises in the sample of the survey (approximately five percent of the three million enterprises), and a significant percentage of them (often around five % within an economic sector) have their values revised. This update may be an effect of two different causes: the lack of recent updating of the register, or a recent and real economic change of the activities of the enterprise. For example, some enterprises are leaving their industrial activities, and develop mainly commercial activities.

Sector-based statistics are important, particularly if considered from a political point of view: for example, the turnover in the manufacturing industry. It has to be noted that this kind of statistic is produced using not one variable but two, one categorical and one quantitative. For example, the total turnover of sector X is

$$\sum_U Turnover(i)1_{APE=X}(i) \tag{1}$$

where $1_{APE=X}(i)$ is the variable indicating whether the enterprise *i* belongs to the sector X (its APE code is equal to *X*).

Using purely administrative data combined with the value of the code available in the business register would lead to serious problems due to lack of updating, and consequently result in biased estimates. Accordingly, the information obtained through the statistical survey has to be used: for the enterprises belonging to the sample we have two values, one not updated, the other one corresponding to the present situation.

As shown in Figure 1, the available material consists of an incomplete rectangular table, with a considerable proportion of the data missing. How can this material be used to produce estimates? For this purpose, two methods have been studied:

- Mass imputation (Kovar and Whitridge 1995), which means imputing values for all variables in the questionnaire of the statistical survey for the nonsampled enterprises (that constitute 95% of the population, even if they are small). For example, the value of the APE code should be imputed taking into account what has been observed in the sample for the changes between classes. This method belongs to the model-based approaches family of:
- Using combined statistical estimates, with a design-based approach.

Studies (Brion 2007) have been conducted on past data showing that the estimates of totals as "total turnover of a sector" obtained through mass imputation are generally biased

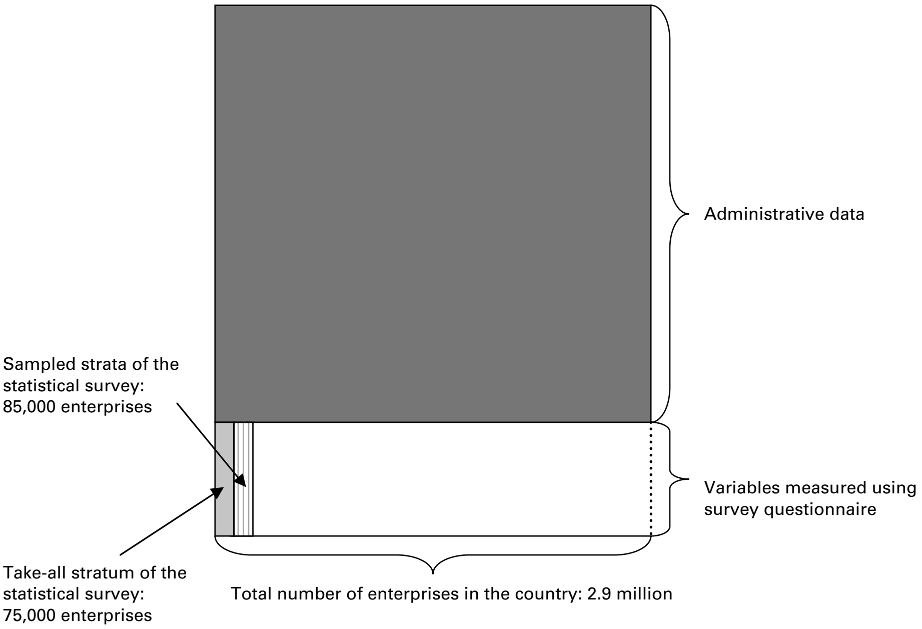
Administrative data
Sampled strata of the statistical survey: 85,000 enterprises
Variables measured using survey questionnaire
Take-all stratum of the statistical survey: 75,000 enterprises
Total number of enterprises in the country: 2.9 million

Fig. 1. ESANE, a multisource device for the French structural business statistics. A schematic representation of ESANE: rows represent variables and columns represent enterprises. The upper part of the figure contains variables obtained through administrative sources, and the lower variables obtained through the statistical survey. The take-all stratum of the survey contains larger enterprises (generally defined as more than 20 employees). The white area dominating the lower part represents unobserved data (since in the sampled strata only 85,000 enterprises are surveyed from the population of almost 3 million units)

(even if the imputation of the classifying code is unbiased), and that the bias may be important in some sectors (more than ten % of the estimated aggregate).

The reason is that one would like to impute the value of the code (qualitative variable) conditionally on other variables (the turnover for example, but all administrative variables – that means hundreds of variables – are in fact concerned, which makes the conditional imputation impossible in practice). On the same subject, Kroese and Renssen (2000) present similar conclusions on the limits of the mass imputation methods.

Now, by implementing ESANE, it was also considered that the question of the APE code (economic classifying of the enterprise) should not be the only one leading to a "comparison" between statistical survey and administrative data. The value of the turnover, and of its breakdown into different activities (for example industrial production, trade, production of services), is asked for in the statistical survey and is also available in the fiscal source. This possibility of "linking" the two sources is seen as playing an important role to give robustness to the system: in cases of important inconsistencies between the two values, survey clerks are asked to check the cause and to propose an arbitrated value. This phase of comparison exercises a quality control of the two sources (administrative and survey), and may reveal some events (for example restructuring) that are not well taken into account in one source.

Having all this material, the question is how to use it in an efficient way. INSEE decided to choose a method based on combined statistical estimates (within the design-based approach). For example, the total of a fiscal variable (turnover, value added, etc.) of a sector X is estimated with the following difference estimator

$$\sum_{U} Yfiscal(i)1_{APEreg=X}(i) + \sum_{s} wi(Ytrue(i)1_{APEsurvey=X}(i) - Yfiscal(i)1_{APEreg=X}(i)) \tag{2}$$

which uses the data coming from the exhaustiveness of the fiscal data and the classification within the register (first term of the Equation (2)), combined with a "correction" resulting from the sample at two levels:

- correction of the classification in the register, using the classification collected through the statistical survey $1_{APEsurvey=X}(i)$, as $1_{APEreg=X}(i)$ is the information available in the business register;
- correction of the "quality" of the administrative – mainly fiscal – data (for example for the variable turnover, but also for variables linked to it, such as goods sales) through the quality control operated on the sample ($Yfiscal(i)$ being the basic value, available for each unit in the administrative source, $Ytrue(i)$ being the value considered as the final value after arbitration, available only for the units in the sample).

The studies conducted (Brion 2007) showed that the results obtained with this method at aggregated levels have a much smaller mean square error than those resulting from mass imputation, and are then more "secure". For example, at the global level of the French trade sector, the square root of the mean square error of the Estimate (2) is half of that obtained with the mass imputation method. Having a comparison at a smaller level of the nomenclature (four digits), it was found that for 13 classes, the mass imputation method shows better results, and for 100 classes the combined statistical estimator has a smaller mean square error.

Those results confirmed the choice of the estimate; again, it has to be noted that mass imputation, in this specific case, would concern 95% of the enterprises.

The final estimators used are, in fact, somewhat more complicated than presented above, since they also take into account some calibration leading to modified values of the final weights (see Brion 2009 for more details). And the treatment of missing data (in both administrative and survey data) makes things more complex than the presentation in this discussion, which is limited to the principles of the method.

It can be noted that the difference estimator presented here is different from GREG estimators, since it uses the richness of the whole administrative data (through its exhaustiveness, for all available administrative data), as GREG estimators would only use the exhaustiveness of a few administrative variables.

However, these combined estimators have some limits: particularly, they do not guarantee always to produce positive values. In some cases, enterprises changing their APE code and having one variable with a large value and/or a big sampling weight may create problems in the second term of Equation (2), leading to negative values. One might think that using some winsorization procedures could help avoid this problem; but since hundreds of variables are potentially concerned, it is practically impossible to implement such a method. Even if this kind of situation (negative estimates) is a drawback for the production of results, it does in fact reveal a problem that would not necessarily appear when using classical methods: in the case of the existence of such units, when a sample without winsorization is used, the estimates would have a very large variance, and when administrative data is used directly with approximate values of the APE code (using the value available in the register), we would have a large bias. In this way, the combined estimates act as a safeguard.

Then, to avoid being faced with a lot of potentially negative estimates for “small categories”, it has been decided to modify the strategy concerning the estimators: at a relatively aggregated level (the three digits level of the nomenclature) the estimates use the Equation (2), and for more detailed levels a breakdown is applied, through percentages obtained *via* the survey (Gros 2012).

To conclude on this point, for the specific problem INSEE has faced, it seems to me that other solutions may exist, specifically using the ideas developed in R. Little’s article. It would be interesting to see some work conducted on evaluating the relative efficiency of these alternatives.

2. An Important Problem for Statisticians in Charge of Production Processes: Data Editing

The control of data is one of the most time-consuming parts of the production process of official statistics. What is more, one may notice that academic literature has focused much more on the question of missing data than on the question of suspicious data.

For some kinds of surveys, especially mail surveys, which represent an important part of business surveys, the activity of data editing (checking the plausibility of the collected values, and having a strategy to decide what to do with them, by modifying them or not) takes a lot of time and energy. The statistician has to work on the definition of appropriate checks (not only for accounting coherence, but also for internal coherence within a questionnaire, for example by studying the plausibility of the ratio of two variables), and on an appropriate strategy to make the manual work of the survey clerks as efficient as possible.

This last question has been studied in many papers during recent years, and some pragmatic approaches – namely selective editing – have been proposed (see, for example Lawrence and McKenzie 2000; Hedlin 2003), relying on score functions quantifying the potential impact of raw data on an estimate. For example, the DIFF function is calculated as:

$$w_i(z_i - y_i) \tag{3}$$

where

w_i is the sampling weight of unit i,
z_i is the value of the raw data for unit i, and
y_i is the value of a predictor, for the considered variable and for unit i (that may be the value of the variable Z of the same unit for the previous year).

The idea is to focus the work of manual checking by clerks on those units whose score value lies above a certain threshold, the other units being treated automatically. As far as I know, little work has been done to try to develop a theoretical approach to this problem, and particularly to introduce some Bayesian formulations, except for the paper by Hesse (2005) that gives a general formulation of the problem, and others papers, such as Buglielli et al. (2010), which uses the Bayesian approach to generate the expected value (the predictor) and also robust aggregates.

Work should be done to extend these developments. From this point of view, having a modeling approach would not be primarily dedicated to estimation purposes, but rather function as a tool for "guiding" the work of survey clerks. Using a Bayesian approach, calibrated with previous surveys for example, might be an interesting way to make the selective editing efficient. The measurement error may be taken into account through a contamination model for example, but the quality of the predictor also has to be modeled, since this quality may be very different from one variable to another, and result in higher or lower efficiency.

This topic offers some interesting challenges for statisticians, partly meeting the concerns presented in Little's article about the total survey error paradigm.

3. References

Brion, P. (2007). Redesigning the French Structural Business Statistics, Using More Administrative Data. Proceedings of the Third International Conference on Establishment Surveys, June 18–21, 2007, Montreal, Canada. CD-ROM: American Statistical Association: pp 533-541.

Brion, P. (2009). L'utilisation combinée de données d'enquêtes et de données administratives pour la production des statistiques structurelles d'entreprises. Paper presented at the Journées de Méthodologie Statistique, INSEE, Paris.

Brion, P. (2011). Esane, le dispositif rénové de production des statistiques structurelles d'entreprises. Courrier des statistiques n°130, INSEE, Paris.

Buglielli, M.T., Di Zio, M., and Guarnera, U. (2010). Use of Contamination Models for Selective Editing. Paper presented at the Q2010 Conference, Helsinki.

Gros, E. (2012). ESANE ou les malheurs de l'estimateur composite. Paper presented at the Journées de Méthodologie Statistique, INSEE, Paris.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics. Journal of Official Statistics, 19, 177–199.

Hesse, C. (2005). Vérification sélective de données quantitatives. Document à lucarne E2005/04, INSEE, Paris.

Kovar, J. and Whitridge, P. (1995). Imputation of Business Survey Data. Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds). Hoboken, NJ: John Wiley.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Kroese, A.H. and Renssen, R.H. (2000). New Applications of Old Weighting Techniques – Constructing a Consistent Set of Estimates Based on Data from Different Sources. Proceedings of the Second International Conference on Establishment Surveys, June 17–21, Buffalo, New York: pp 831-841.

Little, R.J.A. and Smith, P.J. (1987). Editing and Imputation for Quantitative Survey Data. Journal of the American Statistical Association, 82, 58–68.

Received June 2012
Revised September 2012