Survey and Administrative Data Mix in a Business Survey

Silvia Biffignandi•, Leopoldo Nascia*, Alessandro Zeli*

The most important surveys on enterprises in Italy are changing from simple data collection by means of ordinary statistical surveys to a more complex management of business surveys. This implies the recasting of many data sources, primarily administrative data. The objectives and purposes of utilising administrative data are: improving timelines, increasing precision and reducing the statistical burden.
In this paper we approach the problem of improving survey accuracy by the integration of non-response. Innovations in the SCI survey have been introduced over several years and are still in progress. The final enterprises data now come from a complex procedure based on mixing information sources.
A statistical analysis of the gain in data accuracy is useful in understanding how the implemented procedures were successful and identifying new steps that are needed for further improvements. In this paper we analyse the differences between estimates coming from different integration procedures. We propose a methodology for comparing the old and new integration methods and we apply it in evaluating the performances of various approaches and the persistency of results in time (two survey editions are considered: 2005 and 2006). The results of our analysis demonstrate, for the main economic variables of the survey involved in the administrative data integration process, the superiority of the mixed approach with respect to the old approach.

1. Introduction

European countries are setting up methodologies to ensure business survey data quality according to SBS regulation requirements. Many problems arise in implementing new procedures and the improvements obtained should be evaluated. In our paper we focus on the SCI survey[1], which is an annual census regarding firms with 100 or more employees active within the Italian manufacturing industry and services.
In an 'ideal' world for researchers, all companies answer in a short time providing all the information required, fully reflecting true values and the total population is in a 'steady state'.

• Department of Mathematics, Statistics, Informatics and Application, DMSIA, University of Bergamo (Italy),
(research project, ex 60%, 2007, University of Bergamo)
* Istat
[1] It is a mandatory survey in order to match the requirements of Eurostat SBS regulation. It also represents an intermediate data collection for National Accounts estimates. The survey is carried out according to the normative guidelines of the 4th EEC Directive scheme under the Italian national Law No. 69 of 26 March 1990 and the national Legislative Decree No. 127 of 9 April 1991 . The survey collects data concerning profit-and-loss accounts and balance sheets**.** Moreover, information regarding employment, investment, personnel costs and certain regional items is also collected.

Unfortunately the return rate is low (slightly more than 50%) so the validated questionnaires are only a minority (less than 50% of the total population). Editing is necessary to deal with outliers, logical incongruence and with partial non-responses. Survey procedures must deal with mergers and acquisitions and other demographic flows. The main goal of internal SCI production process is the development of a database for the total population including all the single units of analysis, so the integration of non-responses (both total and partial) is required.
In the extended draft of the paper, we present an overview of literature on the use of administrative source experiences in various countries, including problems regarding approaches to non-response. In this shorter draft we focus on our methodology and on its application. In section 2, a description of the integration of economic data for non-respondent enterprises is given. In section 3, the methodology for comparing various integration approaches is described and some results are presented. Section 4 gives some concluding remarks.

2. The integration of non-response procedures into the SCI survey

When carrying out the SCI survey, although data collection is aimed at a universal coverage of all enterprises falling within the established range, there is a non-response problem. Several procedures are used to prevent or integrate missing data.
At the first stage, item non-response is handled. A method of imputation is used so as to replace missing value, such as:

- "a priori" distribution of variables based on sets of information coming from previous surveys,
- accounting links with other variables in the questionnaire.

At the second stage, the unit non-response problem is dealt with. In order to improve data quality and reduce the survey burden a complex procedure is applied which involves mixing survey and administrative sources. Among the available administrative sources, a core of three sources has been chosen for integration for statistical purposes. Each administrative source is characterised by a different degree of coverage of the population of interest and by different sets of variables. Many preliminary studies have been performed for using these administrative sources for statistical purposes (for instance, Monducci, Falorsi, Pallara and Russo, 2003).
The three administrative databases are:

- the National Security data base (its acronym is *INPS*);
- the Chambers of Commerce Financial Statement register. From now on we shall briefly refer to this source with the name BIL;
- the Income Tax Annual Form Register, delivered by the *Agenzia delle entrate*, the agency owned by the Ministry of Finance that handles the management of income tax.
  We shall briefly refer to this source using the name TAX.

The SBS survey mainly used the two latter sources [2]and our analysis takes only those into account.
The BIL data concerns all limited enterprises in Italy. The data set contains about 180 variables from profit-and-loss accounts, balance sheets and part of the explanatory notes.
This dataset is used to benchmark and check data returned from ordinary surveys and to obtain information regarding non-response units.
The TAX data concerns all enterprises, and relates to the annual income tax forms completed by the enterprises. From year 2003 it is possible to use, for statistical purposes, income tax data from all enterprises with a lag of 15 months from the end of the reference year.
Preliminary analyses (Nascia, 2007) showed a substantial similarity between balance sheet and fiscal data, so the use of fiscal data (TAX dataset) in order to acquire information (at least for main profit and loss account items) for non limited non respondent enterprises started for reference year 2005.

[2] INPS data are used only in the production process of SBS preliminary estimates.

The current integration procedure for dealing with non-response combines different approaches and uses the two above-mentioned administrative databases: balance sheets and fiscal data archives.
The first step in the non-response integration process consists of donor integration (Istat 2009). The population of enterprises is stratified according to three variables: economic activity (Istat, 2003), size class and geographical area.
In the second step of the integration procedure, the replacement of the value of the BIL available variables is carried out. These variables are then checked with their totals. The subtotal values not available from BIL estimation are elaborated by means of the values obtained from the donor integration procedure. In particular, if Y_1^D is the donor estimate for the variable Y this value is substituted using the correspondent available BIL value. So: $Y_1^D = Y_1^B$; if the BIL source supplies no data for calculating the total Y_1^B, subtotal data are estimated (details of this procedure are to be found in our extended paper).
This procedure is adopted for integrating all non-respondent companies. Integration procedures substitute the donor values for each variable available in BIL. BIL integration provides values for more than 30 variables.
For those enterprises, mainly unlimited companies, that cannot be matched to the BIL dataset, a third step of the integration procedure is carried out. The third step consists in replacing the donor value with the available TAX data. This procedure is identical to the BIL replacement step described above. The residual share of non-respondents remains completely donor based.
The final database includes 4 broad categories of records (Zeli, 2006): responding firms, firms integrated with balance sheets, firms integrated with fiscal data, firms integrated with donors only. The latter represent a small part of the total database and are not considered relevant, both because they are few in number and because of their negligible weight in terms of economic variables.
Table 1 shows the share of records obtained from respondent enterprises or from those integrated with balance sheets or with other administrative sources for the 2005 and 2006 surveys.

Table 1 . SCI survey – Share of final records by origin – Year 2005-2006 *(Percentage values)*

	Source			
Year	Respondent	Balance sheet	Fiscal data	Donor
2005	46.7	44.9	5.4	3.0
2006	42.1	49.6	4.5	3.8

For the 2005 and 2006 survey the achieved coverage of the target population (using data coming from both survey and administrative sources) is around 96 to 97 per cent. In 2007 a further improvement has been made, since the coverage is 98.2 per cent of the target population.

3. A comparison of the performance of non-response estimation approaches

The aim of this paper is to study: a) the achieved improvement in accuracy connected to the use of a complex integration procedure; b) the bias due to the use of traditional non-respondent integration; c) the relationship between the variation due to the new integration methodologies and classification variables. The analysis is carried out for the 3 most important economic variables collected in the SCI survey and also available in administrative sources: turnover, total costs and personnel costs. These variables are very important in value-added estimation and in calculating of most important average values.
First of all we attempt to demonstrate the relationship between enterprise size and participation in the SCI survey, and how this relationship creates a bias in final estimations. In order to achieve our target we utilise a probit regression between a respondent-non-respondent dummy and the number

of persons employed. The model considers, as dependent variable, the "response" coded as equal to "0" if the questionnaire is not returned and coded as equal to "1" in the case of a response, i.e.

$$\Pr(Y=1) = \Phi(\beta_0 + \beta_1 PE)$$

where PE is the number of persons employed.
The second step is to demonstrate the similarity between respondents and administrative data distributions (mainly BIL) in order to justify the use of the BIL source for non-response integration. In this framework we are also going to demonstrate the difference between donor integrated non-respondent distribution and administrative data integrated non-respondent distribution. At first distributions are compared graphically. Then, in order to implement our analysis, the Fligner-Policello (F-P) test (Fligner and Policello, 1981; Hollander and Wolfe, 1999) is utilised. This test assumes neither normality nor equal variances. It does not even assume that the two distribution have a similar shape. The F-P test is widely used in many fields of study. The F-P test can be interpreted as a test of stochastic equality between two distributions, rejection of the null hypothesis means that the two distribution are different in probability. In the case of rejection of the null hypothesis the sign of FP statistic points out which of the two distributions is dominant: a positive sign means that donor values have an higher probability to take greater values (i.e. overestimate) of given economic variable.
The final step of the study is intended to formalise a relationship between the difference in donor and administrative data integration and the categorical variables of the survey: enterprise size, economic activity and geographical area. This analysis is implemented by the estimation of the following model:

$$_jd_i = a + \beta_1 {}_jSIZE_i + \beta_2 {}_jES_i + \beta_3 {}_jGA_i + \beta_4 {}_jR^{t-1}{}_i + \varepsilon$$

j=1,2 1 = balance sheet integrated units, 2=tax data integrated units
i=1,..,3 1= turnover, 2=total costs, 3=personnel costs.

Where *d* is the logarithm of the absolute difference between the values deriving from the donor integration process and the value deriving from the administrative data integration process, SIZE is a dummy of enterprise size in terms of people employed, ES is a dummy of the economic activity carried out by the enterprise, GA is a dummy of the enterprise location and R^{t-1} is a dummy of the presence of an enterprise response for the previous SCI survey edition.
The last two available editions of the SCI survey (2005 and 2006) are considered in our analysis; the results are similar in both years. The first step is the estimation of the probit model for detecting the influence of size on the probability of response: although low, a relationship between enterprise size and SCI questionnaire response exists and it is a positive one.
The second step is an analysis of distribution equality for three variables: turnover, total costs and personnel costs.
We initially compare the distribution of respondent enterprise and BIL data sources for respondent enterprises. For the variables considered, the distributions of respondents and BIL values do not present significant differences and they are graphically similar (in the graphical analysis the variables are logarithmically transformed).
The comparison between BIL and donor integrated distributions indicates a substantial difference between distributions. Donor distributions are apparently located on the right with respect to those for BIL. The results of the F-P test describe a dominance of donor distributions with respect to the BIL distribution for all variables. As an example of the F-P test we show in Table 2 the results (for the year 2005) related to the comparison between donors approach and BIL source integration.

Table 2 – F-P tests for differences of distributions of main economic variables for donors and BIL source – Year 2005

variable	n	test	1-tailed-p	2-tailed-p
turnover	4721	3.41317	.00032	.0006
total costs	4721	3.10956	.00094	.0019
personnel costs	4721	4.55859	.00000026	.00000051

The positive FP test values suggest that the distributions are statistically different, moreover the donor group has a larger median and the donor distributions dominates the BIL distribution for all variables.
The same test is carried out for the subset of non-respondent enterprises integrated with the TAX source for turnover and total costs. The test verifies a clear difference between donor and TAX source distributions.
As the final step of our analysis, the regression models are implemented to verify the effects of stratification variables on the differences between donor integration values and the BIL integration method. The estimated models show that there is a difference between donor and BIL values that can be explained by the strata determined by stratification variables. As a matter of fact, the coefficients are all quite positive and significant for all variables, so It highlights that the greater the enterprise, the greater the bias using the donor methodology. For geographical area and economic activity the differences regarding the respective benchmarks are significant, although the effects are not as strongly evident as those for SIZE. Whether or not the enterprise is respondent or non-respondent for the previous edition of the survey has no effect on the differences between integration methods.
The estimates for differences between donor integration and TAX data confirms that there is strong evidence for the differential effect of the SIZE on the difference between donor value and the final value taken from TAX data sources. The differential effects with respect to the benchmarks for the other variables are significant, but weaker.

4. Concluding remarks

The analyses and tests proposed in our study confirm that although the donor methodology is the best in the absence of other sources, SCI survey data accuracy is increasing since the administrative source has started to be utilised. In particular, the use of two administrative sources (balance sheets and tax data) is both relevant in correcting for non-response and allows for a relevant gain in total population coverage, while guaranteeing the quality of the estimates.

References

Fligner M.A., Policello, G.E. (1981) *Robust rank procedures for the Behrens-Fisher problem*, Journal of the American Statistical Association 76 (373) pp. 162-168.
Hollander M., Wolfe, DA. (1999) *Nonparametric statistical methods (2nd ed.),* New York: Wiley.
Istat (2009) *Conti economici delle imprese - Anno 2005* Istat Tavole di dati sul web, 6th February.
Monducci, R. et al. (2003*) Prime esperienze sull'utilizzo integrato di fonti statistiche ed amministrative per la produzione di statistiche strutturali sui risultati economici delle imprese,* in Temi di ricerca ed esperienze sull'utilizzo a fini statistici di dati di fonte amministrativa*,* a cura di Falorsi P.D., Pallara A. Russo A., Franco Angeli, Milano.
Nascia L. (2007) *Administrative data in SCI survey data processing* presented at the workshop on SCI data editing Rome 1st February.
Zeli A. (2006) *The use of administrative data in Structural Business Statistics: the Italian experience* presented at the conference "Towards a New Statistical System" Belgrade, 5th July.