A proposal for the treatment of non-response to improve the quality of business census-like data in Italian National Accounts estimations

Puggioni, Augusto
Istat - Italian National Statistical Institute, National Accounts Directorate
Via Balbo 16
00184 Rome, Italy
E-mail: puggioni@istat.it

Sacco, Giuseppe
Istat - Italian National Statistical Institute, National Accounts Directorate
Via Balbo 16
00184 Rome, Italy
E-mail: sacco@istat.it

1.Introduction

The availability of census-like data sources on enterprises is a key requirement to obtain a sound estimate of economic aggregates. The Survey on Economic and financial accounts of large enterprises (SCI), run by Istat - Italian National Statistical Institute, is one of the main sources used in National Accounts (NA), as it collects information on the annual reports of all the big enterprises with at least 100 workers, in manufacture, construction, trade and services. Non-response imputation in the results of the survey comes out to be an issue of paramount importance. Experience has shown that the traditional imputation methods based on the use of administrative data of the same enterprise to impute the main items of the questionnaire and on the use of a "donor" selected among the respondents to impute the sub-items, proved inadequate when used to reconstruct the information for big enterprises in a time series approach. The paper analyzes alternative approaches to impute non-response and provides some comparative outcomes on how the traditional and the new proposed methods impact on the estimation of the annual changes of economic data.

The next paragraph (par.2) describes the main characteristics of the sources used. After a description of the current imputation method adopted for SCI, paragraph 3 illustrates a new method to deal with total non-response. Paragraph 4 presents some results showing the impact on the estimates of the proposed method. Finally, paragraph 5 concludes with some future research perspectives.

2. Business data in Italian National Accounts
2.1 Survey on Economic and financial accounts of large enterprises

SCI is the main source for estimating the economic aggregates of enterprises with at least 100 workers. It is a structural-type survey carried out every year by Istat on a census basis. In order to reduce the entity of the error on the analysis domains for the estimates of the NA, survey microdata are used. The NA has been using the enterprises survey data on a micro-level for some time now (since the general 1988 revision). The need of having microdata comes, in particular, from the fact that the estimation domains are different compared to those established by the surveys: this involves carrying out further editing processes and using different estimators.

The controls on the microdata are both preliminary and integrated in the construction and analysis process of the accuracy of the estimates of the NA aggregates, as, generally, they provide good indications on the entity of errors that influence the estimates. It is important to point out that these controls are specific to

the construction process of the national accounts and do not substitute but only complete and integrate the analysis processes of each statistical survey.

The use of a specific data control and correction method is due to the difference that exists between the concept of quality of the survey and the concept of quality of the NA estimates: the process used to improve the quality of the NA estimates from the survey data can thus differ from that of the survey. Of course, the two aspects tend, though partially, to overlap each other (Puggioni, 2000).

The integration of the sources by the NA allows identifying any possible errors in the microdata, in addition to evaluating the components of non-sampling error (coverage errors, measurement errors, non-response errors).

This census-like survey is addressed to all enterprises with at least 100 workers, which amount to about 11000 units (Istat, 2010). The data collected refer to the enterprises, classified according to their main economic activity and to their "functional units" (a concept quite similar to that of kind of activity unit, KAU, derived from each enterprise by separating the main economic aggregates based on its different lines of production), in order to supply homogeneous data by sector of economic activity. The classification of the economic activities of the enterprises is the 5-digit "ATECO" (five digits Italian classification equivalent for the first 4 digits to the NACE - Rev.1.1 since 2002 and to the NACE - Rev.2 since 2008).

All multi-territorial enterprises are requested to provide data on a territorial basis, too (administrative region/NUTS2) as to the number of workers, compensation of employees and gross capital formation. The regional value added is estimated proportionally to the regional distribution of compensation of employees as indicated by the enterprise.

All the economic activities are surveyed, except for agriculture-zootechnics-hunting and fishing, financial activities, general government and NPISHs activities and activities carried out by households and cohabitations.

In the case of partial non-responses, missing data are filled by means of the remaining valid answers and by means of the relations of balance existing between the questionnaire items. Such analytical imputation is done both automatically and manually by the statisticians who are in charge of the survey.

As regards the total non-response, two integration methods are applied:

- in the case of corporations, data are integrated with the business accounts;

- as to the remaining enterprises, integration of the variables is done through the so called "donor imputation".

The performance of the integration from business accounts is enhanced by the legal nature of the enterprises included in the field of observation, over 90% of which are made up of corporations.

The total amount of a variable for a certain stratum is equal to the sum of all amounts of the respondent enterprises and of the integrated enterprises.

The coverage degree, in terms of number of enterprises, equals, on average, 45-50%, value that by the integration from the balance sheets and other administrative sources, even reaches about 95-98%.

2.2 Business accounts of corporations

The use of business accounts data is a part of Istat project on the use of administrative archives for statistical purposes, which started in last 90es. This archive is submitted to a careful analysis, which highlights its correct use for statistical purposes. The analysis consists on identifying the conceptual reference framework relative to the information object of analysis, on identifying the statistical target population, the survey and analysis units, the characters, the classifications, the updating periods and modalities, as well as on identifying the rules for converting the administrative data into statistical information (Vaccari, 2002).

The enterprises must present their balance sheets to the Chambers of Commerce, which will then make them available and consultable within their network (Infocamere). Business accounts did indeed prove to be adequate to represent many of the variables held in ISTAT questionnaires (Dabbicco, 2001).

As mentioned in paragraph 2.1, these data are used for imputing the total non-response of SCI survey (see Dabbicco, De Gregorio, 2002, as to the methodological details).

3. A proposal for a different imputation procedure to deal with non-response

Data from SCI survey are used for National Accounts estimations according to estimation domains which are different from those provided for by the SBS Regulation (Nace Rev.2 and Size Class – 100-249, 250-499, 500 and more workers in the case of National Accounts estimations, Nave Rev.2 without size class for SBS Regulation estimations). The data are subjected to a procedure which is based on the following steps:

- calculation of the value of production and value added at producer prices, intermediate costs at purchase prices for each microdata item (by firm and by functional unit);

- data quality analysis: any outlier (whether as observed or imputed data, or as classification of economic activity declared or attributed to the firm, if non-responding) are analysed at the micro level, in collaboration with the statisticians who are responsible for SCI survey, to be validated and in some cases corrected;

- aggregation of data on firms, according to the stratification described above.

As to the integration of a non-response, this is carried out by the statisticians in charge of the survey by using imputation methods based on a minimum distance "donor" and, in second place, by statutory accounts, for income statement data, if the data item is available in the Istat database, imputing the items present in the survey with their counterparts in the accounts (the so-called "main items" in the income statement).

The imputation of the questionnaire items, which by definition are not present in the accounts, is still carried out in any case using the "donor" method, while the "donor" is used in a complete way to impute all the items of the questionnaire when the firm to be integrated is not present in the accounts database at all.

In recent years, the distribution of data on firms in terms of respondents and non-respondents has been as follows (table 1):

Table 1 – 2005-2008 SCI Survey

SIZE CLASS	FIRMS	SOURCE_1	SOURCE_2	SOURCE_3
year 2005				
Total	10510	4908	317	5285
100-249	7295	3096	210	3989
250 +	3215	1812	107	1296
year 2006				
Total	10805	4547	427	5831
100-249	7485	2902	303	4280
250 +	3320	1645	124	1551
year 2007				
Total	11166	4936	202	6028
100-249	7748	3105	165	4478
250 +	3418	1831	37	1550
year 2008				
Total	11275	4795	270	6210
100-249	7767	2856	228	4683
250 +	3508	1939	42	1527

Where:

Source_1 : data collected by survey;

Source_2 : data imputed by "donor";

Source_3 : data imputed by administrative data (in large part, business accounts).

To limit the negative effects that the use of the "donor" has on National Accounts domain estimations, where the longitudinal aspect of the information is significant as well as its structural aspect, an imputation method is proposed based on the following rules, given the final surveys for year t and year t-1, a method which can be defined as that of the "annual change donor". Specifically, the integration of the non-response is obtained by splitting the data set relating to year t into two data sets consisting, respectively, of potential donor firms (firms correctly recorded in the survey) and of non-respondent firms. The two data sets thus identified are subsequently stratified according to a number of prominent variables; this makes it possible to operate on subsets of similar firms in terms of economic activity, geographical area and size class.

The same operations of partitioning and stratification are performed in parallel for the archive relating to year t-1; moreover, only the firms present in both years are taken into consideration.

The donor firm is therefore sought within the stratum corresponding to that of the receiving firm.

The choice of donor firm is made through recourse to a suitable distance function which uses information relating to a number of prominent variables (number of workers, economic activity, geographical area).

If several firms at a minimum distance are identified, they are selected random. Finally, the internal items of the questionnaire are reconstructed by applying the changes in the per-capita values in the two years observed in the donor firm.

Being $X_{i,j}^{t}$ the questionnaire item j relating to the non-responding firm I, $P_{d,j}^{t}$ and $P_{d,j}^{t-1}$ the per-capita values observed at times t and t-1 also for variable j relating to firm d chosen as the donor, one obtains:

$$X_{i,j}^{t} = (P_{i,j}^{t-1} + \Delta_{d,j}^{t,t-1} * P_{i,j}^{t-1}) * A_{i}^{t} \quad \text{where} \quad \Delta_{d,j}^{t,t-1} = \frac{P_{d,j}^{t} - P_{d,j}^{t-1}}{P_{d,j}^{t-1}} \quad \text{and} \quad A_{i}^{t} \text{ is the number of workers,}$$

assumed to be known at time t for firm i.

The procedure described is not applied for items inventories and fixed assets; empirical considerations have shown such a wide variability for these variables that it is inadvisable to use information relating to the changes $\Delta_{d,j}^{t,t-1}$ in the per-capita values. For these variables we have

$$X_{i,j}^{t} = P_{d,j}^{t} * A_{i}^{t}$$

just like the current integration phase for SCI non-responses, i.e. by calculating firstly the per-capita values of the "donor" for each questionnaire item and then attributing to the firm to be imputed these per-capita values multiplied by the firm's number of workers or employees, in the case of labour cost items, available from the Statistical Business Register (ASIA).

The data integrated from the "donor" constitute a very small share of the target population in the case of final data (approximately 2%). From a theoretical point of view, therefore, the use of the method described is legitimate; their share is obviously more significant for provisional data, used by national accounting for semi-final estimations of national accounts. In any case at the domain level (Nace 4 digit – classes 100-249, 250-499, 500 and more workers) they may influence the aggregate result also in the final version of the survey.

As to firms integrated with business accounts data, the method described above may be used to reconstruct the items contained in the questionnaire, i.e. not "main item".

In any case, for the sake of reliability, as the percentage of non-responding firms imputed with accounts data is significant, it is not advisable to use this imputation method. As a possible development of this work the following procedure is proposed.

1) Order the firms integrated with accounts data by domain (Nace Rev.2) in descending order in relation to number of workers;

2) for each domain select a number of firms so that their share is not greater than 10-12% of the target population;

3) use the annual change donor, described in the previous method, taking care finally to square the data in relation to the value of the main item imputed by business account.

4. Results

The actual effectiveness of the method proposed for non-responding firms for which accounting data are not available has been evaluated for the 2008 survey data in Nace Rev.2.

In table 2, the current method ("CURR") and the new method ("NEW") turnover per capita 2008-2007 annual changes are presented, for size class and total.

Table 2 – Results - Turnover (thousand euros)

SIZE CLASS	FIRMS 2008	WORKERS 2008	"CURR" TURNOVER 2008	"NEW" TURNOVER 2008	"CURR" PC ANNUAL CHANGE %	"NEW" PC ANNUAL CHANGE %
Total	146	33396	6590867	5278419	56.1	25.0
100-249	111	17252	2831596	2947526	27.3	32.5
250 +	35	16144	3759271	2330893	87.7	16.4

5. Conclusions and future works

This paper presents a new method of imputation for non-response in a census-like business survey. The main innovations that were introduced concern the treatment of non-response in a time series approach. That is very important for National Accounts as underlined above. The use of microdata is, in fact, one of the main characteristics of the construction method of the national accounts used by Istat. This approach makes it possible to use statistical sources more efficiently, while also allowing for a more complete integration.

This work emphasizes, to this purpose, the differences with the current method for non-response used in SCI, designed for different goals (structural data). A good improvement has been recorded in the annual changes of turnover per capita. Anyway, it is important to point out that the improvement of the accuracy of the estimates did not affect the total economy "original" levels of the economic aggregates of the survey, which means that a good level of consistency with the published data has been ascertained.

As a future development, we will attempt to continue along the complete calibration of this model, for firms integrated with business accounts data too.

REFERENCES

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). Statistical matching. Theory and practice. Chichester (UK), John Wiley and Sons.

Dabbicco G. (2001), Utilizzo dei bilanci aziendali civilistici ai fini del soddisfacimento del regolamento UE 58/97 sulle Statistiche Strutturali sulle Imprese (SBS-Structural Business Statistics), unpublished report, Rome, Italy, Istat.

Dabbicco G., De Gregorio C. (2002), L'utilizzo dei dati dei bilanci civilistici per l'integrazione delle mancate

risposte totali alla rilevazione sul Sistema dei Conti delle Imprese (SCI), unpublished report, Rome, Italy, Istat.

Istat (2010), Conti economici delle imprese – Anno 2007.

Puggioni A. (2000), L'analisi di qualità delle stime di contabilità nazionale, Seminario "La nuova contabilità nazionale", ISTAT, Rome, 12-13 January 2000.

Puggioni, A., Sacco, G., (2010), L'utilizzo dell'indagine SCI in Contabilità Nazionale e proposte di nuovi controlli sui dati, unpublished report, Rome, Istat.

Rassler, S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. New York, Springer.

Vaccari C. (2002), Caricamento e analisi dei dati di bilancio formiti dalla soc. Pitagora. Accoppiamento con ASIA, unpublished report, Rome, Italy, Istat.