Exploiting the integration of businesses micro-data sources

Giovanni Seri – Daniela Ichim - Valeria Mastrostefano – Alessandra Nurra
Italian National Statistical Institute (Istat)
National Accounts and Business statistics Department
{seri,ichim,mastrost,nurra}@istat.it

1. Introduction

Recently the Italian National Statistical Institute (Istat) is evolving towards an integrated production system of SBS statistics. In this model, the core of the information content is represented by administrative sources while sample surveys are conducted in order to estimate only not directly available specific sub-populations information. For the majority of enterprises, the core of SBS variables, such as Turnover, Purchases of goods and services or Personnel costs, are often registered in different administrative sources, such as Financial statements, Tax Authority data or social security data. Consequently, the core of SBS variables may be estimated at an extremely refined resolution level. For other variables, complex statistical modeling strategies might be required. The SBS variables obtained through administrative data collection or statistical estimation procedures are registered in an exhaustive archive, called Frame, covering the whole population of enterprises as defined by the SBS Regulation.

The paper describes the analyses performed in order to integrate the Frame main information with data stemming from two structural businesses sample surveys. The main objective has been the production of economic indicators exploiting the interaction between two data sources, a register and a sample survey.

The considered sample surveys are the Community Innovation Survey (CIS) and the Information and Communication Technologies (ICT) survey. It is worth noting that the core variables of each survey are not registered in the Frame. Examples of such variables are binary indicators regarding innovation status, type of innovation activities, use of mobile devices or involvement in e-commerce activities.

Two different statistical approaches were taken into account, a macro and a micro one. The first one applies to aggregated (tabular) data stemming from the linking of the Frame and the chosen survey. The aggregates are then conveniently modified and subjected to the constraint that the marginal distributions independently derived from the two sources are maintained. The latter approach applies to a linked microdata file. Indeed, a 'new' set of weights is calibrated in order to preserve the consistency with the surveys disseminated statistics . Simultaneously, the totals derived from the Frame variables are accounted for. Several calibration strategies are compared in this study.

In the paper we will illustrate the results obtained by implementing the two considered approaches. In the Section 2 a brief description of the data used is given. In Section 3 the implementation of the different methods is described. Section 4 is devoted to comment the results obtained. Some conclusions are drawn in Section 5.

2. Data

The main objective of the project is to develop a strategy for deriving economic performance indicators by combinations of structural and survey information. The economic indicators considered in this work are all registered in Frame, e.g. Value added per person employed, ratio between Value added and Turnover, etc. The spanning variables are defined as pairs of structural information registered in Frame (principal economic activity; size class) and survey indicators. Possible inconsistencies between the data sources were solved by assuming that the true/real information was registered in Frame.

The reference year for the data used in this work is 2012. The registration of a key identifier facilitates the linkage of the data sources. As regards the Frame, the main variables Turnover, Personnel costs, Other costs, Value added, Operating profitability (EBITDA), Number of persons employed, NACE, Region were considered.

ICT and CIS surveys: purposes and indicators

The used survey data concern the ICT survey conducted in the year 2013[1], and the most recent edition of the CIS survey (the 7th Europe-wide CIS) referring to enterprises innovation activities between 2010 and 2012.

As for the ICT survey[2] a set of six indicators, characterizing the enterprises with at least ten persons employed operating in industry and non financial services were defined. Different topics belonging to the ICT questionnaire were included in the analysis through the derivation of composite indicators. Two main criteria were used for the indicators selection. The first criteria is intrinsically represented by the aim of the current project. Indeed, the trivial replication of already disseminated information is out of scope. Secondly, as Frame information is yearly registered and archived, 'core' questions and areas which are observed each year were identified. Consequently, biennial indicators or those belonging to the one-off sections characterizing the dynamic nature of observed ICT phenomena were avoided.

Based on national or international experiences (European Commission, OECD composite indicators related to the following areas of interest were included in this work: downloading speed of Internet connection declared by businesses (e_speed), intensity of use of the network in terms of persons employed using Pc connected to the Internet for work reasons, dematerialization and integration of organizational processes, levels of maturity reached by the company in e-commerce (from those only buying on line to those firms selling and buying on line or having also own website offering opportunities to place on line orders for goods and services). The indicators choice leave open the possibility to update and/or extend their definition in order to better monitor the ICT improvement. Indeed, the classification of maturity levels may be easily changed, the speed or ICT usage classes may be updated, as well as the surveyed technologies (for example from Pc to mobile devices intensity usage) The Community Innovation Survey is one of the major sources of innovation data. Based on a 'subject' approach aimed at identifying the innovative behavior of firms, its main goal is to overcome some drawbacks of the traditional long-established indicators based on the science-push model of innovation (R&D and patents indicators). CIS provides data on a diverse range of ways of innovating and captures forms of 'dark innovation' that don't rely on formal in-house creative activities such as R&D and which are seldom patented. CIS explores as well small-scale innovation or technology adoption of the "off-the-shelf innovators".

In particular, the survey covers innovation activities of the Italian enterprises with at least ten persons employed operative in industry and services and focuses on four macro-typologies of innovation: product, process, organizational and marketing innovation, even if just for the first two categories it collects more detailed information on the expenditures, outcomes, linkages, sources for knowledge and technology transfers, factors hampering and objectives of innovation.

The survey is part of the Eu Innovation Survey (Cis), carried out on a two-year basis (from 2004 onwards) by all the Eu Member States and candidate countries. In order to ensure a sound comparability across countries, the Cis is carried out on the basis of a standard core questionnaire and a harmonized survey methodology developed by Eurostat, in close cooperation with the participating countries. Since 2000, the Cis has become one of the major sources of data for the European Innovation Scoreboard, and it has confirmed by the European Commission one of the flagship initiatives for measuring the performances of the Innovation Union within the Eu2020 strategy.

In this preliminary phase, in the selection of the most suitable indicators we have privileged some complex indicators based on the responses to different nominal level questions, more revealing of firms strategies than simple indicators and best capturing the propensity of the Italian firms to

[1] The reference year should be 2013, but since the survey was conducted in the half of 2013 it is possible to consider the required qualitative information as referred to the end of 2012.

[2] Since 2004, data collection on ICT is based on a European Regulation which ensures that the data are harmonized among Member Countries and in line with strategic European framework for the information society. ICT survey produces indicators for Digital Agenda Scoreboard (one of the seven pillars of the Europe 2020 Strategy) and it is annually implemented to better respond to evolving needs by users and decision makers.

innovate, here defined as the attitude to carry out any kind of innovation activity (product, process organizational and marketing innovation, R&D driven or not) and regardless of whether the activity resulted in the implementation of a commercially successful innovation.

ICT and CIS surveys: methodological framework

Both ICT and CIS are surveys governed by specific European Regulations requiring estimates for given domains of the target population, i.e. enterprises employing at least 10 persons and belonging to certain NACE code[3].. The one stage stratified random sampling (without replacement) is the sampling design of both ICT and CIS surveys. The strata are defined by combining the economic activity (Nace classification), size class (Number of persons employed) and region (Nuts classification) according to the domains of interest. Equal selection probabilities are assigned to enterprises belonging to the same stratum. The sample size in each stratum is mainly defined according to the Bethel procedure (Bethel, 1989) as the minimum sample size ensuring that the coefficient of variation of pre-defined estimates in given given domains does not exceed a given threshold. Estimates are then derived through calibration methodology (Deville, Särndal, 1992; Casciano *et al.*, 2006) to compensate nonresponses and to match known population totals (benchmarks) of selected auxiliary variables (Number of persons employed, Number of enterprises). The latter population totals are computed using the Italian Statistical Business Register (ASIA). According to the time schedule of the surveys, the reference year of the ASIA register is 2011 and 2012 respectively for ICT and CIS. When linking Frame and ICT datasets, due to the different reference years, over 19114 units, about 1400 units cannot be linked. The main reason is given by the changes occurred in the number of persons employed. Consequently, the majority of the non-linked enterprises did not satisfy the criteria defining the survey target population. Additionally, several NACE misclassifications and demographic events caused around 100 ousting of units. As regards the integration of Frame and the CIS survey this kind of problems have a very low impact, as the Frame and the survey sampling frame (the most updated version of the official statistical business register Asia) both refer to the same reference year (2012). Anyway, the linking do not cover the whole CIS target population as the Frame excludes the Financial Services sectors which are considered in.

3. *Methods*

We considered two different approaches called macro and micro integration. In the first case we considered estimates as two way tables involving both Frame and survey variables. The tables were then perturbed by a multiplicative algorithm and by imposing constrains on the marginal row and column totals. In the second approach, through calibration, the sampling weights (or a set of initial weights) were modified in order to achieve numerical consistency between estimates and 'known' population totals.

Following the macro approach we first considered the method known as Balancing (Nicolardi, 1998; AAVV, 2012) where a set of estimates in the form of tabular data stemming from different sources and having some common have to be reconciled in order to achieve consistency on these margins. The method is usually used in the compilation of the National Accounts and it is implemented as a constrained optimization problem[4]. In our purpose, marginal totals were determined from the two different sources while the initial cell values were computed on the linked dataset: the dataset including statistical units belonging to both the sources (the Frame reduced to the theoretical population of the survey and the sample data set of the survey). Unfortunately, the implementation does not impose non negative constraints for the cell values. In our application, the solution diverges to unacceptable solutions (negative frequency counts). Therefore we tested the Iterative Proportional Fitting procedure (IPF[5]). IPF requires as input the two given marginal distributions and an initial set of cell values. IPF iteratively adjusts the cell values to achieve consistency with the marginal row and

[3] Hereafter for Frame we intend the dataset including enterprises belonging to the theoretical population of the considered survey (196186 units for the ICT survey and 160909 units for the CIS survey).

[4] The method is implemented in an R routine developed at Istat.

[5] Implemented in the R package Teaching Sampling available in R.

column totals. The method may be easily implemented. Since it uses a multiplicative algorithm to achieve the consistency with given marginal distribution, there is no risk to obtain inadmissible solutions. IPF may be applied independently in each table. On one side, this feature increases its applicability. On the other side, without further control or constraints, inconsistencies in linked tables are possible. It is worth noting that the marginal distribution of Frame quantitative variables with respect of survey categorical variables cannot be known; therefore it was estimated by means of the corresponding distribution derived from the linked dataset. We report IPF as method A when presenting the results.

As far as the micro integration approach is concerned, different calibration strategies were tested[6] (Deville, Särndal, 1992; AAVV, 2012; Leadership Group SAM, 2003). First we applied the calibration strategy used by the survey. Indeed, the population totals of the variables Number of persons employed and Number of enterprises for given domains were derived from the Frame. Then, these totals were used as known population totals when calibrating the weights corresponding to the linked dataset.

In order to achieve consistency on the productivity indicator Value added per person employed, the second strategy, named method B, consists in enriching the set of auxiliary variables by Value added. In order to guarantee the convergence of the calibration algorithm, the geographical information was removed from the list of variables defining the estimation domains.

When combining the Frame and ICT information, an additional method C was implemented. Instead of using the linked dataset, the ICT survey dataset may be directly used, by ignoring its original calibrated weights. In this case, The Frame plays twice the role of secondary source. Firstly, the balance sheet information may be integrated for common businesses. Secondly, Frame may be used for computing the known population totals. Although we achieved numerical consistency with the known totals, the main drawback of this strategy is that the theoretical population depends on the ???. Consequently, the method C may generate incoherencies. That's why this strategy was applied only to the ICT survey. Moreover, for the CIS survey the linked data set and the full sample survey dataset are extremely similar and significant differences between the two set of weights resulting from the two strategies were not observed.

Finally, we tested a strategy, called D, where the known totals computed using the Frame auxiliary variables were combined with the ICT estimates derived for a categorical variable (e.g. the number of enterprises performing ICT sector activities or not). The idea behind the strategy was to simultaneously calibrate on known population totals for the variables involved in the computation of the productivity indicator and to be consistent with the selected estimates of the ICT indicator. Subsequently, by means of Consistent Repeated Weighting – CRW (AAVV, 2012), different ICT selected indicators were added. In general, in our tests, when the ICT estimates used as known population totals were null or very small, the algorithm did not converge.

4. Some results

The integration strategies illustrated in the previous section were applied for different economic indicators and for different combinations of spanning variables. A selection of the results obtained is reported in Tables 1 to 6.

In Table 1 the Value added per person employed (VA/PE) is reported for the subpopulations of enterprises defined by cross-classification of the NACE categories and the downloading speed of the broadband Internet connection; the latter variable is called e_speed. The NACE categories were grouped in "inside" and "outside" the ICT sector while the categories of the binary ICT indicator "e_speed" were defined using a threshold equal to 10 Mbit/sec. The shown results allow for the comparison of the four strategies: (A) IPF; (B) calibration of the 'linked dataset' using known totals derived by the Frame; (C) calibration of the 'survey dataset' using known totals derived by the Frame and (D) calibration of the linked dataset using known totals derived from Frame and ICT, respectively.

[6] The generalized software ReGenesees was used. The software has been developed at Istat (http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/regenesees).

It is worth noting that each third column is constant, proving the convergence to the known values of the marginal distribution (differences reported for method D are within the admissible error range).

In Table 2, through the relative differences of the values reported in Table 1, the IPF method is compared with calibration approaches (B, C and D), while the method C is compared with the method B. Similar conclusions may be drawn for other comparisons were performed for different cross-tabulations involving more detailed NACE levels and other ICT indicators.

In Table 3 the percentage of Value added out Turnover is reported. The cross-classifying variables of the Table 1 are used. In this case the economic indicator involves the Turnover information which was not considered as auxiliary variable during the stratification and calibration processes.

Similarly, we present in Table 4 the values of Value added per Person Employed (VA/PE) computed on the CIS data. We consider a different classification of the NACE code in six categories defined by the Pavitt taxonomy and the CIS binary indicator "PPI" identifying the 'enterprises carrying out product or process innovation'. As stated before, only two methods are considered: (A) IPF and (B) calibration of the 'linked dataset' on known totals derived from the Frame. Table 5 reports the comparison of this methods through relative differences of the values given in Table 4. In Table 6 the percentage of Value added over Turnover is reported in for the same cross-classification of Table 4.

As expected, the analysis of Table 1 indicates a positive correlation between higher Internet connection speed and potential for greater use of the technologies and higher efficiency which induces greater value added per person employed values than companies connecting at speeds below 10 Mbit/sec.

Similarly on the base of the CIS-Frame data we can convey that there is a positive correlation between higher values of the economic performance indicators (value added per person employed and ratio between value added and turnover) and the implementation of innovation activities.

In both cases, the results presented here are partial. Other aspects related to the compliance of data obtained with the expertise of the phenomena detected require further study. However, the proposed methods lead to sensible conclusions both from the mathematical and subject-matter points of view.

5. Conclusions and further work

In this work two different approaches were taken into account in order to exploit the integration of two sources of business data. The datasets considered in this study are represented by an exhaustive archive, called Frame, supplying the main balance sheet variables for the whole population of enterprises as defined by the SBS Regulation and a sample survey dataset adding thematic variables not registered in the Frame. A macrointegration and a microintegration approaches were tested. A general comparison of the two strategies is a difficult task as it should depend on the available data and on the aim of the integration project. In any case, subject-matter experts should always be involved in the quality analysis of each integration project. The Balancing methodology can simultaneously deal with a set of tables, while IPF deals with a single table, independently on any other information. In both cases nothing can be stated about precision of the estimates. Microintegration through calibration takes into account the detail of domain of estimates that allow for convergences. In particular, the calibration of the linked dataset may be preferred as the direct calibration of the survey dataset reduces the importance of the Frame. Moreover, the calibration on known totals stemming from disseminated estimates does not always converge. Further studies on calibration methodology will be done considering different sets of auxiliary variables to produce alternative indicators. Moreover we could test also the possibility to define different sets of weights for different target indicators.

Another way to exploit the information supplied by the Frame in favor of sample survey is to consider the Frame as the business register (instead of ASIA) to draw samples using economic variables not elsewhere available.

Finally we should mention the possibility of simultaneously integrating the two sample surveys and the Frame. This objective will be pursued by statistical matching technique (D'Orazio *et al.*, 2006). At the moment we can stated that only the population of enterprises in Industry (NACE divisions 10 to 33) is the 'common universe' allowing for statistical matching analysis on CIS and ICT data.

Bibliography

AAVV (2015). Il sistema informativo frame SBS per la stima del conto economico delle imprese basato sull'uso di fonti amministrative. *Numero Speciale Rivista di Statistica Ufficiale* Istat (in corso di pubblicazione).

D'Orazio M., Di Zio M., and Scanu M. (2006). *Statistical Matching, Theory and Practice*. Wiley, New York.

Bethel, J. (1989) "Sample allocation in multivariate surveys". Survey methodology, 15 (1989): 47-57.

Casciano, C., e P.D. Falorsi e S. Filiberti e A. Pavone e G. Siesto (2006). "Principi e metodi per il calcolo delle stime finali e la presentazione sintetica degli errori di campionamento nell'ambito delle rilevazioni strutturali sulle imprese". Rivista di Statistica Ufficiale, n. 1 (2006): 67-102.

Deville, J.C., Särndal, C.E. (1992). "Calibration Estimators in Survey Sampling". Journal of the American Statistical Association, 87 (1992): 376-382.

AAVV (2012). Essnet on Data Integraion Final Reports http://www.essnet- portal.eu/project-information/data-integration

Eurostat (2012). Final Report ESSnet on Linking of Microdata on ICT Usage, (link: ec.europa.eu/eurostat/documents/341889/725524/2010-2012-ICT-IMPACT-2012-Final-report.pdf).

Leadership Group SAM, 2003, Handbook on Social Accounting Matrices and Labour Accounts, Population and Social Conditions 3/2003/E/N23.

Nicolardi, V. (1998), "Un sistema di bilanciamento per matrici contabili di grandi dimensioni', (A balancing method for big accounting matrices), Istat, Quaderni di ricerca, n. 4, 1998.

Statistics Sweden (2010). Yearbook on Productivity 2010, Are ICT Users More Innovative? An analysis of ICT-enabled Innovation in OECD Firms by Vincenzo Spiezia OECD *(link: (http://www.scb.se/statistik/_publikationer/OV9999_2010A01_BR_X76BR1001.pdf)*.

VA/PE	IPF (method A)			Linked datasets (method B)			Survey' dataset (method C)			Table (method D)		
	e_speed			e_speed			e_speed			e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	49082	62457	55065	48905	62976	55065	48529	63435	55065	50520	61363	55056
Inside ICT sector	52313	123658	104070	51801	122924	104070	50770	124932	104070	54442	125040	104265
Tot_e_speed	49168	67433	57600	48977	68006	57600	48588	68483	57600	50625	66725	57600

Table 1. Value added per person employed for the ICT and non-ICT economic activities and e_speed values: comparison of the methods A, B, C and D

Rel Diff VA/PE	(A-B)/A			(A-C)/C			(A-D)/D			(B-C)/B		
	e_speed			e_speed			e_speed			e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	0,4	-0,8	0,0	1,1	-1,6	0,0	-2,9	1,8	0,0	0,8	-0,7	0,0
Inside ICT sector	1,0	0,6	0,0	3,0	-1,0	0,0	-4,1	-1,1	-0,2	2,0	-1,6	0,0
Tot_e_speed	0,4	-0,9	0,0	1,2	-1,6	0,0	-3,0	1,0	0,0	0,8	-0,7	0,0

Table 2. Relative differences (%) Value added per persons employed for the ICT and non-ICT sectors and values of e_speed: calibration methods B, C and D compared to the method IPF (A) and of method C with respect to B

VA/TURNOVER	IPF (method A)			Linked datasets (method B)			Survey' dataset (method C)			Table (method D)		
	e_speed			e_speed			e_speed			e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	20,9	19,6	20,2	21,4	19,9	20,6	21,2	20,4	20,8	21,4	19,4	20,4
Inside ICT sector	30,9	43,5	41,1	31,1	43,6	41,4	30,5	41,2	39,3	28,9	42,9	39,9
Tot_e_speed	21,1	21,4	21,2	21,6	21,7	21,6	21,4	22,1	21,8	21,5	21,2	21,4

Table 3. Value added out Turnover (%) for the ICT and non-ICT sectors and values of e_speed: comparison of the method A, B, C and D

VA/PE	IPF (method A)				Linked datasets (method B)		
	PPI				PPI		
PAVITT	0	1	Tot_PPI		0	1	Tot_PPI
Not elsewhere classified	66945	110452	81831		67515	112120	81831
High-technology	89509	88624	88837		90627	88231	88837
Medium-high-technology	54347	71570	67341		56533	70933	67341
Medium-low-technology	50065	61042	56180		50603	60703	56180
Low-technology	41953	61195	52800		43984	59747	52800
Knowledge-intensive services	64292	114504	95853		65103	115239	95853
Lessknowledge-intensive services	47237	58403	51877		47879	58302	51877
Tot_PPI	52489	73223	63332		53423	73000	63332

Table 4. Value added per person employed for Pavitt categories and values of PPI: comparison of methods A and B

Rel Diff VA/PE	(A-B)/A		
	PPI		
PAVITT	0	1	Tot_PPI
Not elsewhere classified	-0,9	-1,5	0,0
High-technology	-1,2	0,4	0,0
Medium-high-technology	-4,0	0,9	0,0
Medium-low-technology	-1,1	0,6	0,0
Low-technology	-4,8	2,4	0,0
Knowledge-intensive services	-1,3	-0,6	0,0
Lessknowledge-intensive services	-1,4	0,2	0,0
Tot_PPI	-0,9	-1,5	0,0

Table 5. Relative differences (%) Value added per persons employed for Pavitt categories and values of PPI: calibration methods (B) compared to the method IPF (A)

VA/TURNOVER	IPF (method A)				Linked datasets (method B)		
	PPI				PPI		
PAVITT	0	1	Tot_PPI		0	1	Tot_PPI
Not elsewhere classified	20,9	13,5	16,7		19,2	12,5	15,6
High-technology	25,4	33,3	31,0		23,1	29,7	27,7
Medium-high-technology	21,3	23,2	22,8		21,0	22,0	21,8
Medium-low-technology	16,9	19,7	18,5		15,7	19,0	17,5
Low-technology	20,5	21,3	21,0		21,7	21,6	21,6
Knowledge-intensive services	34,9	39,4	38,2		32,6	41,3	38,6
Lessknowledge-intensive services	14,2	16,4	15,2		15,4	17,1	16,1
Tot_PPI	18,0	20,6	19,5		18,1	20,3	19,3

Table 6. Value added out Turnover (%) for Pavitt categories and values of PPI: comparison of methods A and B