Exploiting the integration of businesses micro-data sources

Giovanni Seri, Daniela Ichim, Valeria Mastrostefano, Alessandra Nurra [1]

Abstract[2]

The new Statistical information system for estimating structural economic variables on business accounts (Turnover, Value Added, ...) based on the primary use of integrated administrative/fiscal data, "complemented" with survey data, FRAME SBS, has been released in 2012-2013. FRAME-SBS is by now the pillar of the new system of economic statistics in Italy. As further development, the exploitation of new opportunities of economic analysis by the integration of Frame SBS with other sources of data stemming from sample surveys has been particularly promoted. The paper reports on the work done to define a methodological approach for the production of economic indicators involving variables stemming both from Frame SBS and from two sample surveys: Community Innovation Survey (CIS) and Information and Communication Technologies Survey (ICT).

Keywords: SBS, Information and Communication Technologies Survey (ICT), Community Innovation Survey (CIS), data integration, balancing, calibration.

1 Researcher (Istat), e-mail: {seri,ichim,mastrost,nurra}@istat.it.

2 A version of this paper was presented at the 4th European Establishment Statistics Workshop held in Poznan (07-09 September 2015) and is available on the workshop website (http://enbes.wikispaces.com). The authors are grateful to Orietta Luzi who contributed with useful suggestions. Although the article is the result of a joint work, paragraphs 1, 2.1, 2.3, 3.2, 5 has been drafted by Giovanni Seri, paragraph 2.2 by Alessandra Nurra and Valeria Matrostefano, paragraph 3.1 by Daniela Ichim and paragraph 4 by Giovanni Seri and Daniela Ichim. The views expressed in this paper are solely those of the authors and do not involve the responsibility of their Insitutions.

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. 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 examined, 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 micro approach applies to a linked microdata file obtained by merging a survey dataset and the Frame archive. Practically, 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 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 on the results obtained. Some conclusions are drawn in Section 5.

2. Data description

2.1 Frame SBS

In Italy, SBS estimation has been traditionally based on data collected through two direct annual surveys: the sample survey on Small and Medium Enterprises (SME) (enterprises with less than 99 persons employed, about 4.3 million of units as reference target population), and the total survey on Large Enterprises (LE, about 11,000 enterprises representing the census target population for all enterprises with 100 or more persons employed). Both surveys collect information according to EU harmonised statistical definitions on profit and loss accounts, as well as on employment, investments etc. in the industrial, construction, trade and non-financial services sectors.

The increasing stability, timeliness, coverage and accuracy of firm-level information available in some administrative sources on businesses' economic accounts has made it possible to develop a new SBS estimation system mainly based on the direct use of administrative data as primary source of information, integrated with the SME and LE survey data. Firm-level data for the main economic aggregates are directly obtained from the integrated sources, as they cover about 95% of the whole target population. As a consequence, aggregates of the most important SBS variables can be determined at an extremely high precision level, while for other SBS variables, more complex statistical modeling strategies might be required. The resulting statistical data warehouse covering the whole SBS target population and variables is named Frame.

It should be pointed out that each combined source actually covers different, possibly partially overlapping, subpopulations of enterprises, and that some sources provide data on a, possibly partially overlapping, set of variables. The overlapping information has been used for assessing the quality of input data and harmonizing classifications and definitions with SBS concepts described by the SBS regulation. Specific analyses have been devoted to manage inconsistencies among data from different sources. The registration of a key identifier facilitates the linkage of the data sources. (in each administrative archive enterprises are uniquely identified and classified based on a complex procedure performed at the Business Register construction stage).

The reference year for the data used in this work is 2012. The main objective of the project is to develop a strategy for deriving economic performance indicators by combinations of register 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.

2.2 ICT and CIS surveys: purposes and indicators

The used survey data concern the ICT survey conducted in the year 2013[3], 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[4] 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 criterion 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 their 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

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

4 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.

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 CIS 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 been confirmed by the European Commission as 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 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.

2.3 ICT and CIS surveys: methodological framework

Both ICT and CIS are surveys ruled by specific European Regulations requiring estimates for given domains of the target population, i.e. enterprises employing at least 10 persons and belonging to given NACE codes[5]. The sampling design of both ICT and CIS surveys is one-stage stratified random sampling. 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

5 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).

Bethel procedure (Bethel, 1989) as the minimum sample size ensuring that the coefficient of variation of estimates in predefined domains does not exceed a given threshold. Estimates are then derived through calibration methodology (Deville, Särndal, 1992; Casciano *et al.*, 2006) to compensate nonresponse and to match known population totals (benchmarks) of selected auxiliary variables (Number of persons employed, Number of enterprises). The 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, around 1400 units of the 19114 units cannot be linked. The main reason is given by the changes in the number of persons employed. Consequently, the majority of the non-linked enterprises did not satisfy the criteria defining the survey target population (at least 10 persons employed). Additionally, several NACE misclassifications and demographic events caused around 100 ousting of units. As regards the integration of Frame and the CIS survey these kinds 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 does not cover the whole CIS target population that includes the Financial Services sectors which are considered in Frame.

3. Methods

3.1 Macro-integration

Following a macro integration approach we considered estimates as two way tables involving both Frame and survey variables. Particularly, the spanning variables are structural information registered in Frame as NACE or size class combined with a survey indicator. The cells contain aggregations of an economic variable/indicator registered in Frame as Value added or the Number person employed. The tables were then modified by a multiplicative algorithm and by imposing constraints on the marginal row and column 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 marginal 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[6]. For our purpose, marginal totals were

6 The method is implemented in an R routine developed at Istat.

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 target population of the survey and the sample data set of the survey). Unfortunately, the implementation does not impose non-negativity 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[7]). 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 column totals. The method may be easily implemented. Since it uses a multiplicative algorithm to achieve the consistency with a 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 to 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.

3.2 Micro-integration

In the micro-integration 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.

Different calibration strategies were tested[8] (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 to present the results, 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. Moreover, the calibration process is generally set to achieve a-priori defined lower and upper bounds for weights.

7 Implemented in the R package Teaching Sampling available in R.

8 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).

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 through its original set of calibrated weights assigned to all the units in the survey sample. In this case, the Frame plays twice the role of secondary source: firstly, the Frame balance sheet information may be integrated for common units; 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 target population selected by the Frame does not include the whole sample and this can represent a sort of incoherency. For the CIS survey the linked data set and the full sample survey dataset are extremely similar (no significant differences between the two sets of weights resulting from the two strategies were observed). That's why this strategy was applied only to the ICT survey.

Finally, we tested a strategy, called D in the results, 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[9] 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 (and maybe published) 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 zero or very small, the algorithm did not converge.

4. 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 4.1 to 4.6.

In Table 4.1 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

9 ICT sector in NACE Rev. 2 (based on the 2006 OECD definition) is defined by the following economic activities: 261, 262, 263, 264, 268, 465, 582, 61, 62, 631, 951 (https://ec.europa.eu/eurostat/cache/metadata/en/isoc_se_esms.htm). It is possible to distinguish ICT Manufacturing activities (261, 262, 263, 264, 268) and ICT services activities (465, 582, 61, 62, 631, 951).

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 estimates, respectively. 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 4.2, through the relative differences of the values reported in Table 4.1, 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 that were performed for different cross-tabulations involving more detailed NACE levels and other ICT indicators.

In Table 4.3 the percentage of Value added out Turnover is reported. The cross-classifying variables of the Table 4.1 are used. In this case the economic indicator involves the Turnover information which was not considered as auxiliary data during the stratification and calibration processes.

Table 4.1 – Value added per person employed for ICT and non-ICT economic activities and e_speed values: comparison of methods A, B, C and D

VA/PE	IPF (method A) e_speed			Linked datasets (method B) e_speed			Survey' dataset (method C) e_speed			Table (method D) e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	49,082	62,457	55,065	48,905	62,976	55,065	48,529	63,435	55,065	50,520	61,363	55,056
Inside ICT sector	52,313	123,658	104,070	51,801	122,924	104,070	50,770	124,932	104,070	54,442	125,040	104,265
Tot_e_ speed	49,168	67,433	57,600	48,977	68,006	57,600	48,588	68,483	57,600	50,625	66,725	57,600

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Table 4.2 – Relative differences (%) Value added per persons employed for ICT and non-ICT sectors and values of e_speed: calibration methods B, C and D compared to method IPF (A) and of method C with respect to B

Rel Diff VA/PE	(A-B)/A e_speed			(A-C)/C e_speed			(A-D)/D e_speed			(B-C)/B 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

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Table 4.3 – Value added out Turnover (%) for ICT and non-ICT sectors and values of e_speed: comparison of method A, B, C and D

VA/TURNOVER	IPF (method A) e_speed			Linked datasets (method B) e_speed			Survey' dataset (method C) e_speed			Table (method D) 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

Source: Authors' calculation on ICT and FRAME data - reference year 2012

Similarly, we present in Table 4.4 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 Eurostat/OECD technological intensity classification[10] combined with the CIS binary indicator "PPI" identifying the 'enterprises carrying out product or process innovation'. As stated before, only two methods are considered for the CIS survey: (A) IPF and (B) calibration of the 'linked dataset' on known totals derived from the Frame. Table 4.5 reports the comparison of these methods through relative differences of the values given in Table 4.4. In Table 4.6 the percentage of Value added over Turnover is reported for the same cross-classification of Table 4.4.

Table 4.4 – Value added per person employed for technological intensity categories and values of PPI: comparison of methods A and B

VA/PE	IPF (method A) PPI			Linked datasets (method B) PPI		
PAVITT	0	1	Tot_PPI	0	1	Tot_PPI
Not elsewhere classified	66,945	110,452	81,831	67,515	112,120	55,065
High-technology	89,509	88,624	88,837	90,627	88,231	
Medium-high-technology	54,347	71,570	67,341	56,533	70,933	
Medium-low-technology	50,065	61,042	56,180	50,603	60,703	
Low-technology	41,953	61,195	52,800	43,984	59,747	
Knowledge-intensive services	64,292	114,504	95,853	65,103	115,239	
Lessknowledge-intensive services	47,237	58,403	51,877	47,879	58,302	104,070
Tot_PPI	52,489	73,223	63,332	53,423	73,000	57,600

Source: Authors' calculation on CIS and FRAME data - reference year 2012

10 https://www.oecd.org/sti/ind/48350231.pdf
https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:High-tech_classification_of_manufacturing_industries

Table 4.5 – Relative differences (%) Value added per persons employed for technological intensity categories and values of PPI: calibration methods (B) compared to method IPF (A)

VA/TURNOVER	(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

Source: Authors' calculation on CIS and FRAME data - reference year 2012

Table 4.6 – Value added over Turnover (%) for technological intensity categories and values of PPI: comparison of methods A and B

Rel Diff VA/PE	(A-B)/A PPI			Linked datasets (method B) PPI		
PAVITT	0	1	Tot_PPI	0	1	Tot_PPI
Not elsewhere classified	2.9	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

Source: Authors' calculation on CIS and FRAME data - reference year 2012

As expected, the analysis of Table 4.1 shows greater value added per person employed values for companies with higher Internet connection speed than companies connecting at speeds below 10 Mbit/sec confirming a positive correlation between potential for greater use of the technologies and higher economic efficiency.

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 or efficiency of the

estimators 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 future work

In this work we dealt with methodologies suitable to exploit the potential of data 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 approach 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. As for the macrointegration approach methods referring to reconciliation of tabular data were tested: Balancing that can simultaneously deal with a set of tables, and IPF that deals with a single table, independently on any other information. Balancing generated inadmissible solutions. On the other side, IPF was not deemed flexible enough to be applied on a large set of tables. Consequently, they were not further investigated in our case study. Eliminating these drawbacks, microintegration was preferred. Microintegration was implemented through calibration taking into account the detail of domains of estimates that allow for convergence. In particular, the calibration of the linked dataset, i.e. method B, 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 did not always achieve convergence. 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 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 techniques (D'Orazio et al., 2006). The 'common universe' allowing for statistical matching analysis on CIS and ICT data is only the Industry (NACE divisions 10 to 33), thus excluding the service sector. When applying this method, Frame would represent the overlapping information linking the two surveys. The methods to be used to best exploit the results of a statistical matching procedure will be studied.

References

AA.VV. 2012. *Essnet on Data Integration Final Reports.* https://ec.europa.eu/eurostat/cros/content/data-integration_en.

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

Casciano, C., P.D. Falorsi, S. Filiberti, A. Pavone, and 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. Roma: Istat.

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

D'Orazio M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching, Theory and Practice*. Hoboken, New Jersey, U.S.: Wiley.

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*). Quaderni di ricerca, N. 4, 1998. Roma: Istat.

Spiezia, V. 2011. Are ICT Users More Innovative? An analysis of ICT-enabled Innovation in OECD Firms. OECD *Journal: Economic Studies*, Vol. 2011/1, http://www.oecd.org/economy/growth/are%20ict%20users%20more%20innovative.pdf.