Technology, innovation and economic performances: a microdata integration strategy proposal

De Santis Stefano[1], Di Cecco Davide[2], Di Stefano Roberta[3]

[1]ISTAT, Rome, Italy

[2]UnitelmaSapienza, Rome, Italy

[3]Sapienza, Rome, Italy

Abstract

The relationship between technology use, innovation, and economic performance is a crucial topic in the economic literature. Our work proposes different integration strategies of two business surveys targeting the same enterprises population, one aiming at the use of ICT the other at innovation (CIS). Firstly, a simple record linkage is performed which uses CIS as pivot survey, retrieving the respondents present in the same three-yearly period in the corresponding ICTSs (harmonisation of reference period). Secondly, a Statistical Matching approach is used, to impute the missing information. The results of the matching are evaluated in terms of the preserved marginal distribution of the variables imputed in the synthetic dataset. All alternative proposals represent zero-burden solutions to provide an integrated dataset of microdata, which constitutes an important source of information for research/policy purposes and to support official statistics in exploring complex causal effects and gain new insights into these fundamental economic phenomena.

Keywords: ICT, Innovation, Data Integration, Statistical Matching

1. Introduction

The relationship between technology use, innovation, and economic performance is of significant interest and topicality in the economic literature (Di Vaio et al 2021, Gërguri-Rashiti et al 2017, Gomez et al 2017), as well as for policymakers for its implications and positive spillovers on the entire economic system. For obvious reasons of statistical burden, the simultaneous collection of these variables has always constituted a major problem. However, two distinct business surveys are currently available, one aimed at the study of innovation Community Innovation Survey - CIS, the other at the use of technology (Information and Communication Technologies Survey - ICTS). We illustrate the analysis on the Istat surveys on Italian enterprises, however, the two surveys are conducted on the basis of defining criteria and methodologies harmonized by Eurostat. Therefore, our proposal is potentially extendable to all countries involved in their implementation. Both surveys target all enterprises with more than 10 employees, classified in one of the following NACE sectors: C, D, E, F, G, H, I, J, L, M (with the exception of division 75), N, and group 95.1 of section S. The surveys are complete for enterprises with more than 250 employees.

The ICTS is conducted on an annual basis, designed for about 35,000 enterprises. It collects information about the degree of use of information and communication technologies (e.g., the Internet, broadband, websites, social media, cloud computing), as well as the impact of these technologies in relationships with customers, suppliers, and e-commerce. The ICTS represents the main source for the Digital Scoreboard, which is employed by the European Commission to assess the advancement of Europe's digital economy, and for the construction of the DESI indicator (The Digital Economy & Society Index).

The CIS is conducted on a three-years basis, and is designed for about 39,000 enterprises. It is aimed at collecting information on innovation processes in industry and service enterprises. In particular, the survey collects information about new or significantly improved goods or services (product innovations) and new or significantly improved processes, logistics or distribution methods (process innovations), as well as about organizational and marketing innovation. It provides information regarding the size and sector in which innovating firms operate, the expenditures incurred in introducing innovations (including expenditures on research and development), the innovation objectives and their impact on economic performance, the public funding for innovation and cooperative agreements, the factors hindering innovative activity, and the propensity to patent or use other modes of intellectual property protection. CIS results are employed by the European Commission to monitor the level of innovation and competitiveness across the European Union, to develop indicators on science and technology utilized into the European Innovation Scoreboard, and in analysing EU countries' research policies and their effect on economy.

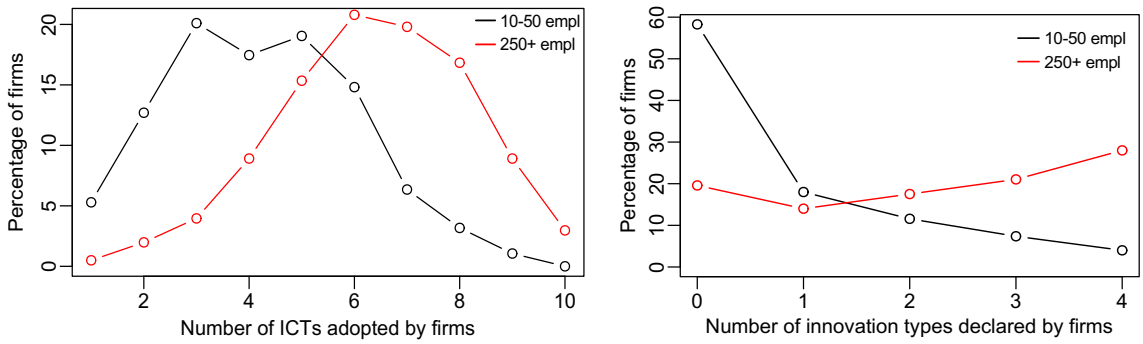
The complexity of the interrelationships between investment in innovation and IT, and all the structural variables of enterprises in different industrial sectors, makes it necessary to consider a joint analysis which would certainly benefit from the availability of microdata with integrated information at the enterprise level. The purpose of this paper is to compare some possible integration procedure to provide a longitudinal dataset with all relevant information without additional burden for the respondents.

2. Data Integration

Given the relevance of the topic, previous attempts at data integration have been made. The common proposed solution (see, e.g., Gierten et al 2021) is to consider for the analysis just the units in common in the most recent year of the three years reference period of CIS. In this approach, the effort focuses on the sampling aspect, i.e., on recalibrating the sampling weights to try to recover the representativeness of the obtained intersection. This approach, in its simplicity, favours the most recent information collected, but suffers in terms of

representativity. In fact, the intersection between the two surveys is only about 15% of the sampled enterprises (about 5,000). More importantly, the intersection is unbalanced towards large enterprises. In fact, the tendency to refrain from including small enterprises in multiple surveys simultaneously to reduce the burden, and the sampling schemes adopted (which include all enterprises with over 250 employees), leads to an overrepresentation of large enterprises. This imbalance, among other issues, leads to an overestimation of the overall tendency towards innovation and the use of technology (see Figure 1).

Figure 1: Differences in innovation and ICT use for class of enterprises



We propose a series of incremental contributions for data integration, with the aim of improving different aspects of estimation efficiency with respect to the simple method mentioned above. Each of these proposals corresponds to advantages, but also imply additional assumptions whose limitations must be carefully weighed against the intended use of the integrated database. Ultimately, the choice between different solutions must be made with respect to the cognitive purposes that have been set.

A) The first proposal for data integration involves utilising the temporal scope of the CIS, which investigates any innovative behaviours carried out in the previous three years (e.g., CIS 2018 investigates innovative behaviours from 2016-18). The proposal is to integrate the CIS with the corresponding three ICTS waves (2016-2017-2018 in the example). The most recent data will be given priority in cases where multiple ICTS interviews have been conducted with the same responding unit.

This approach allows for a notable increase in the intersection coverage, with approximately 50% of units interviewed included. It also facilitates the recovery of data from smaller size classes of enterprises, which were severely underrepresented in the first naive approach. However, this approach does come with certain assumptions/limitations: 1) The freshness of the information is susceptible to change in the interval between the actual ICTS interview and the end of the reference period. This is particularly relevant for firms interviewed in the first

wave (2016 in the example), as the information may not be as up-to-date as it could be. 2) It is necessary to impute some questions, where the ICTS questionnaire provides for temporal alternation of the questions that are submitted to enterprises in each ICTS wave. 3) We implicitly assume that ICTS investments are not reversible. Fortunately, this does appear reasonable, as an investment done in a given year is unlikely to be dismantled in relation to the sunk costs already incurred.

B) As a next step, we explore the possibility of utilising the entire union of the two surveys in a given year. A deterministic record linkage procedure was employed to combine the surveys and the Italian business registers, which presented no difficulties. For the reference year 2016, we have 5,684 enterprises in common, 20,092 that are present only in the CIS and 13,543 only in the ICTS. The idea is to impute the missing information of the units covered by just one survey with statistical matching methods. This procedure allows for a further increase in the sample size, essentially, we have a consistent dataset for each given year, at the cost of the introduction of synthetic data.

C) Finally, we consider an imputation procedure which takes advantage of the enlarged intersection, as we constructed it in the first point. In this way, we exploit a more substantial pooled part to take into account the eventual higher order relationships between the variables to be imputed and the common structural variables.

For the sake of simplicity, this work has limited its focus on the imputation of a select number of the numerous variables included in the surveys. In particular, four dichotomous variables were considered from the CIS, which asked respondents whether there had been innovation in the product, process, marketing or organisational fields. Similarly, 12 variables were considered from the ICTS, which are necessary for the construction of the Digital Intensity Index (DII). The DII is one of key performance indicators on digital performance, and is also used in the construction of the DESI we mentioned in the Introduction. It is used to categorise the enterprises into four classes of "very low", "low", "high" and "very high" digital intensity.

2.2 Statistical Matching

Let V_{ICTS} and V_{CIS} be the variables of interest available only in ICTS and CIS respectively, and let X be the subsample of units in common to both surveys. We aim at imputing the missing variables for each firm not in X. The ideal approach to imputation would be to select and train a multivariate model for the joint distributions of all variables over the complete subset X, and project it over the missing part (see, e.g., Schafer 1997). That is, we would select the variables (V_X, say) more significantly associated with both V_{ICTS} and V_{CIS}, and then impute the missing information according to the conditional distributions $P(V_{ICTS}|V_X, V_{CIS})$ and $P(V_{CIS}|V_X, V_{ICTS})$ in such a way as to maintain the more significative higher order interactions among all variables.

As previously stated, the subset X is not representative of the entire population, as large enterprises (>250 employees) are overrepresented. These limitations restrict the use of such direct imputation methods. We then resort to statistical matching (SM) techniques and compare various approaches which make use of the classic conditional independence assumption (CIA). That is, if V_X denotes the matching variables, the assumption:

$$P(V_{CIS}, V_{ICTS}|V_X) = P(V_{ICTS}|V_X)P(V_{CIS}|V_X)$$

We compared a non-parametric and a semi parametric SM approach to construct the synthetic microdata: Hot deck Nearest Neighbour Donor (NND) and Predictive mean matching (PMM).

By linking the surveys with the business registers, we get access to a large number of structural variables available for all units. Each such variable can potentially be used as matching variable in the SM procedure, without any problems of harmonization between the two samples. An exploratory analysis is conducted to identify the variables that are most significantly related with variables V_{ICTS} and V_{CIS}. Note that we could theoretically utilise two different set of matching variables, (or two different models in PMM), for the imputation of V_{ICTS} and V_{CIS}. However, sector specificity (multi-digit NACE codes) and the variables "number of employees" and "turnover" were found to have the highest association with both the variables on technology use and on innovation, and were thus selected.

In the NND procedure the chosen matching variables are utilized to evaluate the similarities between units, which was measured by the Mahalanobis distance. In the PMM procedure we define a model $V_{CIS} \sim V_{X'}$ and a model $V_{ICTS} \sim V_{X''}$, where $V_{X'}$ and $V_{X''}$ in our case coincide. The models are estimated over the non-missing subsets, and the estimated parameters are used to obtain predicted values for the response variables both in the complete and in the missing parts. The distance between donors and recipients is defined over the predicted values.

In both approaches, for each recipient unit to be imputed, the donor is chosen at random from a set of closest units (five in our application). In order to account for the surveys sampling design, we utilize the stratification variables, in addition to the matching variables, to define the donor sets. That is, the donors research is limited to certain design domains, and, in addition, is sampled with probability proportional to its weight (units with larger weights will have a higher chance of being selected) (Andridge and Little, 2010). Note that this randomized step makes the use of multiple imputation (not explored here) an immediate option. The two approaches lead to similar results. In the next section we present those based on the NND. The number of times the same donors are utilized is an important metric of the SM procedure, as an excessive reuse can affect the variability of the results. Despite the imposed restriction on the donor set for each recipient, and the use of weights, that favour the sampling of units precisely in the

areas with the highest number of missing, 75% of the donors were used only once, with an overall average of 1.9 times.

To assess the CIA, we conducted an analysis on the relationship between V_{ICTS} and V_{CIS}, utilising Poisson generalized linear models. The results demonstrated, perhaps unsurprisingly, that the positive association between the variables remained statistically significant even after accounting for NACE classification, number of employees, and revenue. In particular, enterprises categorised as "high" in terms of digital intensity usage were found to be associated with all types of innovations in many subpopulations. This result was particularly evident among enterprises in the class 0-49 employees which have the lowest propensity to innovate and DII, but also present the strongest association between the two dimensions. Relaxing the CIA appears crucial as the criticality of the assumption is most pronounced for the very class of units which requires more imputation. Then the linkage results based on the three-year CIS were employed to enhance the size of X, and facilitate the recovery of smaller firms. In addition, adopted the approach presented in Singh et al. (1993) to exploit this information in an SM procedure. All processing was done in *R* using the *StatMatch* package.

2.3 Evaluating the results

The matching results are assessed in terms of the similarity of the marginal distributions of the imputed variables in the synthetic and observed datasets.

Table 1 Average number of firms by innovation profile (%) and digital intensity index (mean values).

Year	*N*	*Firms (weighted)*	*Product innovation*	*Process innovation*	*Organisational innovation*	*Marketing innovation*	*Digital intensity index*
			Mean	*Mean*	*Mean*	*Mean*	*Mean*
				Full samples			
2012	18.697	163.347	24,47	25,58	31,25	29,56	34,24
2014	17.532	152.997	20,67	21,08	22,49	22,45	33,03
2016	21.127	157.826	26,69	26,96	27,11	23,78	33,44
				Last Reference year matching			
2012	3.375	163.347	24,48	25,85	29,43	29,95	36,33
2014	4.707	152.997	18,85	21,46	20,25	24,20	34,42
2016	5.335	157.826	25,73	24,50	26,95	22,40	35,14
				All Reference years matching			
2012	6.294	163.347	22,37	25,14	31,92	28,93	35,64
2014	9.592	152.997	21,79	22,65	22,97	22,45	35,96
2016	10.408	157.826	27,96	27,43	26,59	23,14	35,01
				Statistical matching			
2012	18.697	163.347	24,47	25,58	31,25	29,56	34,45
2014	17.532	152.997	20,67	21,08	22,49	22,45	33,43
2016	21.127	157.826	26,69	26,96	27,11	23,78	33,54

Source: Elaboration on ICT and CIS survey data and CIS-ICT synthetic data.

The estimates deriving from the survey are taken as benchmark and compared for each of the proposals. The average difference between the ICT and CIS estimates for each type of innovation (product, process, marketing, organizational) amounts on average to 0.05274 at the subsection level (i.e. an average difference of 5.2% in percentage terms). Table 1 presents the values of the estimates regarding the entire CIS and ICT sample. The differences, in terms of totals, are obviously much lower with a good preservation of the original distributions.

3. Remarks and conclusions

The analyses conducted on the basis of the real context of the data from the period 2010-2016 yielded positive results, given the close link between the phenomena of innovation and digitalisation. This underlines the effective usefulness of linkage strategies (deterministic and probabilistic) with the aim of creating an integrated dataset, according to the different micro/macro scenarios described in the document. The validity of the integration results is assessed by their consistency with the estimates of the two surveys, which constitute the benchmark against which to measure the quality of the integration and the usability of the imputed microdata. The evolutionary nature of the two phenomena (which tend to change significantly over time) suggests two important pieces of evidence. The first is thematic: in-depth knowledge and the advice of thematic experts are necessary to allow the phenomenon to be appropriately framed. The second consequence is methodological. Since the main objective of integration is the joint study of the variables detected separately in the two investigations, the developments must also take into account the evolution of the phenomena over time. The necessity of employing distinct methodologies for the utilisation of disparate datasets, contingent upon the intended purpose (macro or micro), is reiterated. This is evident at both the cognitive and methodological levels. At the cognitive level, this entails the estimation or micro-founded analysis of causal relationships. At the methodological level, this encompasses the comparison of statistical matching methodologies at the macro level and the comparison of matching results with those derived from the use of imputation techniques at the micro level.

References

Andridge, R.R., and Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review*, 78, 40–64.

Di Vaio, A., Palladino, R., Pezzi, A., Kalisz, D. E. (2021). The role of digital innovation in knowledge management systems: A systematic literature review. *Journal of business research*, 123, 220-231.

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

Gërguri-Rashiti, S., Ramadani, V., Abazi-Alili, H., Dana, L. P., Ratten, V. (2017). ICT, innovation and firm performance: the transition economies context. Thunderbird International Business Review, 59(1), 93-102.

Gierten, D., Viete, S., Andres, R., Niebel, T. (2021). Firms going digital: Tapping into the potential of data for innovation. OECD Digital Economy Papers, No. 320, OECD Publishing, Paris, https://doi.org/10.1787/ee8340c1-en.

Gómez, J., Salazar, I., Vargas, P. (2017). Does information technology improve open innovation performance? An examination of manufacturers in Spain. *Information Systems Research*, 28(3), 661-675.

Ha, L. T. (2022). Effects of digitalization on financialization: Empirical evidence from European countries. Technology in Society, 68(C). https://doi.org/10.1016/j.techsoc.2021.101851

Renssen, R.H. (1998). Use of Statistical Matching Techniques in Calibration Estimation. Survey Methodology, 24, pp. 171–183

Sarndal, C.E. and Lundstrom, S. (2005). Estimation in Surveys with Nonresponse. Wiley, Chichester.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. CRC press

Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993). "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption". Survey Methodology, 19, 59–79.