Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey

Leandro D'Aurizio[1] *and Giuseppina Papadia*[2]

We examine two sources of bias for the Bank of Italy's panel business survey of Industrial and Services Firms:

   1) the bias caused by panel attrition; and

   2) the bias created by delays in the distributional data on the reference population, needed for computing the survey weights.

   As for the first source of bias, the estimates strongly dependent on big firms' values are less affected by panel attrition than those representing firms' average behavior, independent of their sizes. Positive economic results make it easier to enroll new firms in the survey, in order to replace firms dropping out because of bad economic performances. However, the economic results of new entrances become more aligned to those of the population, once they enter the sample.

   A very different result emerges for the second source of bias, since, when the population size is highly variable, the information delays produce a bias for the estimates influenced by the contribution of great firms, but the effect is negligible for the estimates not dependent on firm size.

Key words: Business surveys; panel samples; panel attrition; administrative data; auxiliary information.

1. Introduction

Business surveys are often conducted by using a panel sample, with estimates that can be affected by panel attrition. Since these estimates are representative of a reference population by means of survey weights, they can be biased because the weights do not fully take into account the evolution of the reference population. Our article aims to evaluate these two sources of bias by relying on auxiliary information from administrative sources. We will understand the effects of panel attrition by using an integrated archive, matching survey data with financial-statements indicators, available also for the years when the panel units are absent from the sample. We will assess the effects of the second source of bias by fully exploiting the population information.

   Measuring ratios is a typical utilization of business surveys. They can be either ratios between a variable at time t and the same variable at time $t - 1$, in order to measure its relative change over time, or they can be ratios between two different variables. For them,

[1] Italian Authority for the Supervision of the Insurance Sector (IVASS), Research and Data Management Directorate, via del Quirinale 21, 00187 Rome, Italy. Email: leandro.daurizio@ivass.it.
[2] Bank of Italy, Economics and Statistics Department, via Nazionale 91, 00181 Rome, Italy. Email: giuseppina.papadia@bancaditalia.it.

two classes of estimators can be computed: the first one comprises what we define as simple estimates, representative of firms' average behavior, not influenced by the largest units in the sample. The second class includes what we define as aggregate estimates, which tend to be heavily influenced by the big enterprises belonging to the sample, especially if they are over-represented, as in the case of Neyman's samples.

We focus on the Bank of Italy's Survey of Industrial and Service-sector firms (*Invind*, from now on) conducted every year with a panel of around 4,000 enterprises (Bank of Italy 2014), representative of the population of industrial and nonfinancial service-sector firms with at least 20 employees.

The first source of bias we examine arises from firms participating in a survey edition and dropping out from the following one, without that being planned in advance. The term normally used for this phenomenon is panel attrition (Martin et al. 2001). We analyze the attrition effects by using an archive of financial-statement data for the whole reference population. It provides indicators for all the panel units for every survey year, even if they are missing from the sample in several survey editions. We can therefore measure whether, for each survey edition, new entrants and dropouts are different and the main determinants of the propensity to enter and leave the sample.

The second source of bias we assess is caused by the delays of the distributional data on the reference population, required to compute the survey weights. The effect of this lag can be measured for the least recent survey editions, for which complete information on the population is available. We can therefore assess how much the bias of the usual estimates derives from out-of-date population information.

We briefly anticipate the main results of our article. The effects of panel attrition on the aggregate estimates are small, since their values depend on the data of the largest companies, which tend to participate in the survey more regularly than smaller firms. The official estimates, regularly analyzed in the Bank of Italy's reports are of this kind. On the other hand, the smaller firms tend to participate in the survey more erratically and accordingly this fact makes it necessary to carefully interpret the simple estimates. On the contrary, the delays in the updating of the reference population are a source of bias for the aggregate estimates, when they are weighted to be representative of the whole reference population. More precisely, they tend to be biased when the population size is highly unstable, with the bias virtually disappearing for the simple estimates.

The article is organized as follows. Section 2 introduces panel attrition. Section 3 describes the specific features of the panel attrition found in *Invind*. Section 4 describes the indexes we use to assess the effect of panel attrition. Section 5 explains the data integration process. Section 6 evaluates the effects of panel attrition effects on *Invind* and Section 7 analyzes the consequences of the delayed updating of the distribution of the reference population. Section 8 sums up the main results and proposes some solutions to manage the problems highlighted in the article.

2. Panel Attrition: An Overview

Panel surveys use the same sample units for repeated survey occasions. This choice enables researchers to understand transition patterns (Fabbris 1989). An obvious operational advantage of panels over repeated independent samples is that the sample is

selected once before the first survey edition. In most cases, the original panel undergoes an attrition process after the first survey occasion and its composition accordingly changes over time. The statistical literature has widely explored the cases of panel attrition created by some sample units either refusing to participate in the following occasions or exiting the reference population. The units that leave the panel after a given survey occasion are routinely replaced by other ones, that either have been in the panel in previous waves and reenter in this occasion, as replacements of other units, or they are enrolled in the panel for the first time (some of them may even be new entries into the reference population).

If the attrition process decreases the sample size, the standard error of the survey estimates automatically increases. The estimates remain unbiased if the attrition is completely random (Little and Rubin 2002). On the contrary, if the attrition depends on some of the variables of interest for the survey, it becomes a source of bias.

There are some solutions to attenuate these drawbacks. Typically, the units leaving the sample are replaced by others, with observable characteristics (the same used in the survey design) quite similar to those of the replaced units. A more complex solution is to set up panel rotations that periodically discard a part of the panel and make place for new units, that can also realign the sample to the changes in the reference population occurred since the creation of the panel (Trivellato 1999). Rotating panels also spread the response burden and therefore indirectly reduce the attrition caused by it (Ardilly and Lavallée 2007). The method is widely applied in surveys of individuals and households.

These measures maintain the initial sample size and keep the precision of the estimates close to the planned levels. If the hypothesis of attrition totally at random is violated, the replacements of the units leaving the panel cannot totally eliminate the bias, since the units leaving and entering the panel may differ according to characteristics not included in the survey design. The bias arises when such characteristics are correlated with the variables of interest.

Among the recent contributions to the statistical literature on panel attrition, the utilization of many techniques for imputing data, missing because of attrition, can provide useful clues regarding the direction of the bias. For instance, Black et al. (2007) use an array of imputation methods, ranging from simple mean imputation to more complex Bayesian resampling techniques, to reconstruct the missing values for a sample of UK data on car traffic. They measure the variability of the results obtained, in order to assess the bias from missing data and its influence on the estimates.

Deng et al. (2013) use the waves of the Survey on Income and Program Participation, regularly conducted on a representative panel sample of US households, and propose to use a series of refreshment samples, composed by new, randomly selected, respondents, as an auxiliary external source. The new respondents answer the same questionnaire used for the main sample. The differences between the answers of the two respondent sets are used to correct for the bias.

Auxiliary information can also be used as a correction factor of the original weighting system, in order to offset the effects of panel attrition. For example, Afonso (2015) corrects the weights to compensate for the missing data created by attrition in a panel of bank microdata that follows the latest trends of the net interest margin of the banking systems of the 15 major countries of the euro area. The revised weights are used in the

context of Generalized Methods of Moments estimation procedures to produce consistent estimates better aligned to the predictions of economic theory.

3. Panel Attrition in the Invind Survey

We concentrate on the attrition observed between two consecutive waves of the *Invind* survey relative to the years $t − 1$ and t, that we indicate with $I_{t−1}$ and I_t (we do not study the effects of attrition on the longitudinal estimates based on the panel). Some firms participate in $I_{t−1}$, but not in I_t (we refer to them as "dropouts") and are accordingly replaced by other firms that did not participate in $I_{t−1}$ (we call them "new entrants"). We use the term "stayers" to indicate the firms participating in both $I_{t−1}$, and I_t. We use these three terms looking at just two consecutive panel waves, without taking into account what happens in the other ones. We differ here from the standard utilization of these three terms in the literature, indicating the units that respectively drop out of the panel definitively, enter it for the first time, or regularly participate in all the survey editions.

In the single cross-sections, the sample is representative of the cross-sectional population with the help of replacement rules that substitute every dropout with a new entrant, having its head office in the same Italian region, together with economic activity and number of employees as close as possible to those of the dropout.

We consider all the waves available from 2002 until 2013. The wave relative to 2002 was the first with the current reference population (composed by the firms with at least 20 employees belonging to the sectors of nonconstruction industry and nonfinancial private services) and the current sample size of 4,000–4,200 firms (Bank of Italy 2005).

The interviews relative to the wave for year t take place in the first four to five months of the following year $t + 1$. The survey collects the values for the main variables of interest (employment, turnover and investment) for the years $t − 1$, t and $t + 1$ (this last value is a forecast for the current year). The changes for the year t relative to $t − 1$ and for the year $t + 1$ relative to t are accordingly computed by using only one survey wave. If extraordinary events (such as mergers, acquisitions or splits) modify the structure of a panel firm between $t − 1$ and $t + 1$, its data cannot be directly used to compute the average variations and require a special treatment. The data are included in the estimates only if they refer to a set of plants and workers fully comparable over the three years $t − 1$, t and $t + 1$. The comparability is obtained either by anticipating the extraordinary events at the beginning of $t − 1$, or by postponing them at the end of $t + 1$.

Invind weights are cross-sectional. For every survey edition, they make the sample representative of the reference population within strata formed by the combinations of six class sizes (in terms of average number of employees: 20–49, 50–99, 100–199, 200–499, 500–999, 1000–4999, 5000 and over) and eleven sectors of economic activity. A successive post-stratification makes the sample representative of the population also at the geographical level: there are 48 post-strata: north-west, north-east, center, south and islands, referred to the firm headquarters' location, combined with size classes and aggregate economic sectors (Table S1 in Supplementary material). A unique set of survey weights is used for each survey edition, relative to the year t.

Many firms are dropouts or new entrants during the years 2002–2013 (Table 1). On average, 20% of the units in a wave drop out from the following one and are replaced

Table 1. Invind survey: yearly attrition rates (2002–2013).

Year	Total number of firms in the sample	Stayers — Firms participating in the year's wave and still present in the following wave		Dropouts — Firms participating in the year's wave and exiting the following wave		Distribution of the firms exiting the following wave (%)					New entrants — Firms participating in the year's wave and not present in the previous wave		of which: firms participating in the past waves	
		Per cent	Number	Per cent	Number	Firms still active with 20 employees and over	Firms still active with less than 20 employees	Firms involved in mergers, acquisitions, contributions, transfers and splits	Firms no longer active	Total	Per cent	Number	Per cent	Number
2002	3,969	78.2	3,105	21.8	864	69.9	7.1	2.9	20.1	100.0	–	–	–	–
2003	4,135	78.4	3,240	21.6	895	74.2	9.3	2.0	14.5	100.0	24.9	1,030	2.6	109
2004	4,226	80.0	3,382	20.0	844	74.1	7.2	2.1	16.6	100.0	23.3	986	3.4	142
2005	4,386	79.8	3,498	20.2	888	77.5	8.2	2.0	12.3	100.0	22.9	1,004	4.4	195
2006	4,252	80.3	3,415	19.7	837	75.5	9.0	1.6	14.0	100.0	17.7	754	4.7	199
2007	4,063	79.2	3,217	20.8	846	74.6	10.3	1.7	13.5	100.0	15.9	648	5.3	214
2008	3,952	80.3	3,175	19.7	777	78.2	8.1	2.3	11.3	100.0	18.6	735	6.8	270
2009	3,921	80.6	3,162	19.4	759	78.0	7.9	2.0	12.1	100.0	19.0	746	7.4	291
2010	3,937	82.5	3,248	17.5	689	78.4	8.9	0.4	12.3	100.0	19.7	775	8.0	313
2011	4,120	82.4	3,396	17.6	724	77.6	7.3	1.7	13.4	100.0	21.2	872	8.8	361
2012	4,213	82.1	3,460	17.9	753	74.1	8.1	2.0	15.8	100.0	19.4	817	8.6	362
2013	4,215	–	–	–	–	–	–	–	–	–	17.9	755	11.8	498
Total		**80.3**		**19.7**		**75.6**	**8.3**	**1.9**	**14.2**	**100.0**	**20.0**		**6.5**	
of which:														
Non-construction industry		81.1		18.9							19.0			
Non-financial private services		78.3		21.7							23.0			
Average firm size (*number of employees*)		**390**		**222**							**303**			

Source: Bank of Italy's Invind business survey.

by a slightly higher number of units (the sample size increased by 6.5% in the time span examined). An average 75% of the dropouts still belong to the reference population of firms with at least 20 employees, 14% are no longer operational and 10% become not eligible for further participation, either because their workforce drops below the 20-employee threshold or, more rarely, because of events such as mergers, acquisitions, and so on. The stayers' average size is 390 employees, higher than that of the new entrants (303), that is in turn higher than dropouts' average size (222). The last two columns of Table 1 show that, on average, a third of the new entrants participated in past waves before I_{t-1}. The shares of dropouts and new entrants are quite stable throughout the period examined. The economic crisis that began in 2009 and a revision of the survey operations in the years 2006–2008 did not significantly alter this pattern.

Looking at the firms' economic sectors, the attrition is stronger for firms of the nonfinancial private services (Table 1, last row) that became a part of the reference population only since 2002. The interplay of three factors explains this result:

1. it takes time to create a stable panel participation, because the questionnaire is difficult to complete,
2. service-sector firms tend to outsource most of the budgeting and accounting tasks needed to complete essential parts of the questionnaire and this discourages their regular participation in the survey,
3. firms' transformations (by mergers, acquisitions, contributions, transfers and splits) naturally decrease the propensity to participate in business surveys and tend to affect service-sector firms more frequently (Bank of Italy 2015).

4. Simple and Aggregate Indexes

An array of variables can be measured over a firm, either dimensional (such as profits, turnover, investment and number of employees) or dimensionless (e.g., ratios like Return on Assets and Return on Equity, indicated with the acronyms ROA and ROE from now on, extensively used to evaluate firms' performances) and both can be summarized through indexes.

For both kinds of variables, we can follow how their averages vary over the years. Let us indicate with i a unit of the sample of size n_t. The indicators can take two different forms, which we define as simple and aggregate indexes. For both forms, they can be either weighted with the cross-sectional survey weights or can be left unweighted.

The units contribute equally to the simple indexes (Table 2, lower part), regardless of their different sizes, while their contribution to the aggregate indexes (Table 2, upper part) takes place according to their relative sizes. It can be easily shown that the relative size is the value of the variable at $t - 1$ divided by its total (we report below the expression for the unweighted aggregate index):

$$\frac{\sum_{i=1}^{n_t} y_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1}} = \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \frac{y_{i,t-1}}{\sum_{i=1}^{n_t} y_{i,t-1}} \tag{1a}$$

Table 2. Key performance indicators of firms.

| | Aggregate indexes | |
	Dimensional variables[a]	Dimensionless ratios[b]
Unweighted	$\dfrac{\sum_{i=1}^{n_t} y_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1}}$	$\dfrac{1}{\sum_{i=1}^{n_t} z_{i,t}} \sum_{i=1}^{n_t} x_{i,t} z_{i,t}$
Weighted	$\dfrac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}}$	$\dfrac{1}{\sum_{i=1}^{n_t} z_{i,t} w_{i,t}} \sum_{i=1}^{n_t} x_{i,t} z_{i,t} w_{i,t}$

| | Simple indexes | |
	Dimensional variables[a]	Dimensionless ratios[b]
Unweighted	$\dfrac{1}{n_t} \sum_{i=1}^{n_t} \dfrac{y_{i,t}}{y_{i,t-1}}$	$\dfrac{1}{n_t} \sum_{i=1}^{n_t} x_{i,t}$
Weighted	$\dfrac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} \dfrac{y_{i,t}}{y_{i,t-1}} w_{i,t}$	$\dfrac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} x_{i,t} w_{i,t}$

(a) E.g.: turnover and profit. – (b) E.g.: ROA and ROE.

Equation (1a) can be easily generalized to the weighted aggregate index in the following way:

$$\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} = \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} \tag{1b}$$

For a dimensionless ratio $x_{i,t}$ the scale factor is a variable $z_{i,t}$ positively correlated with firm size (generally number of employees or turnover), divided by its total. We will use the turnover for our computations.

5. Using External Sources to Measure Invind Panel Attrition

5.1. The Main Issues

The high attrition levels shown in Section 3 require to assess whether, for every cross-section, the economic performances of the firms entering the cross-section, but not present in the previous one (new entrants) and those of the firms absent in the following cross-section (dropouts) are different. We measure the economic performances in terms of turnover changes, profits changes, ROA and ROE, using the indexes defined in Section 4.

It is also relevant to study the propensity to enter or leave a cross-section, with the first propensity requiring firm-level data for new entrants also for the year prior to their entrance, not directly available from the survey.

We therefore face two problems requiring a data integration process:

1) only turnover is collected in the survey, but not profits, ROA and ROE;
2) data should also be available for the years when the units are absent from the sample.

5.2. The Data Integration Process

5.2.1. The New Archive

Since all the panel firms are limited companies, apart from a negligible number of partnerships (less than 0.1% on average), we integrate the *Invind* data with the *Cerved* archive, a data warehouse for all the Italian limited companies' financial statements filed since 1993. We use the firm VAT number, available on both sources, to exactly match the survey data with the corresponding financial figures. The matching fails whenever the VAT number is missing in *Invind* or it does not find, when it is present, a corresponding VAT in *Cerved* because of errors in one or both sources.

5.2.2. The Matching Quality

A first clue of the quality of the matching is the high percentage of matched units (on average higher than 90%, Table 3), that remains stable (as shown in Figure 1) within the categories of the variables used in the stratification and post-stratification steps. This latter result implies that the survey design also keeps under control the bias caused by using a sample smaller than the original one.

For the average number of employees and the turnover levels, found in both sources, we also examine the individual differences between the two corresponding values from the two sources. If we indicate with $inv_{j,i,t}$ and $cer_{j,i,t}$ the values derived respectively from *Invind* and *Cerved,* relative to the *j*-th variable for the *i*-th matched unit in the year *t*, the size of the absolute difference $\left|inv_{j,i,t} - cer_{j,i,t}\right|$ would depend too much on the unit of measure. We eliminate this effect by using a standardized absolute difference $sad_{j,i,t}$, expressed as:

$$sad_{j,i,t} = \frac{\left|inv_{j,i,t} - cer_{j,i,t}\right|}{\dfrac{inv_{j,i,t} + cer_{j,i,t}}{2}} 100 \qquad (2)$$

Table 3.   *Percentage of firms in Invind annual surveys matched with financial-statement archives (2002–2013).*

	Nonconstruction industry firms	Nonfinancial private services firms	Yearly total
2002	87.8	89.6	88.2
2003	87.6	89.1	88.0
2004	88.6	91.8	89.4
2005	92.4	93.0	92.5
2006	93.3	92.6	93.1
2007	93.1	93.4	93.2
2008	92.1	94.2	92.7
2009	93.2	93.5	93.2
2010	93.4	94.8	93.8
2011	93.8	95.1	94.2
2012	94.3	94.6	94.4
2013	90.3	88.9	89.9
Total	**91.6**	**92.6**	**91.9**

Source: Bank of Italy's Invind business survey and Cerved archive.
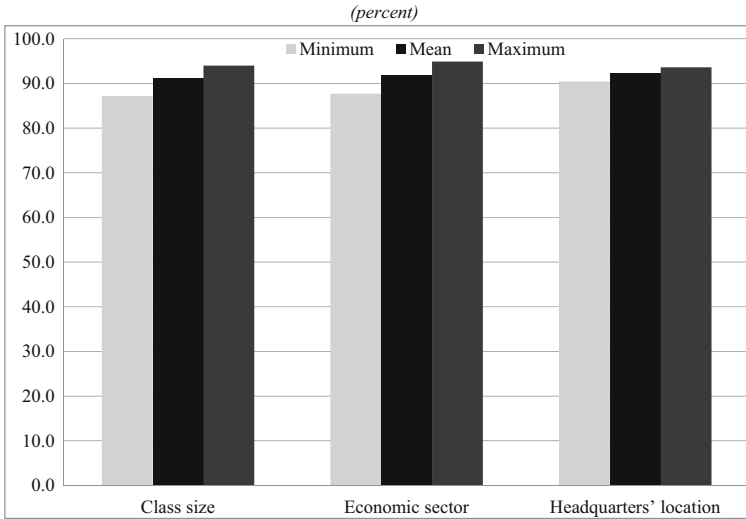
Fig. 1.   *Minimum, mean and maximum of the average percentage of matched Invind units for all the categories of class size, economic activity and headquarters' location. Source: Bank of Italy's Invind business survey and Cerved archive.*

We discard from the analysis the firms affected by structural changes (5.4% for the years 2002–2013), for which the indicator is naturally big, because in such cases the *Invind* figures are adjusted and not comparable with the corresponding ones from *Cerved* (see Section 3 for details on the adjustment method).

The percentage of observations with same values in the two sources are 23.8% (Table 4) for the average number of employees and 11.0% respectively for the turnover levels. The size of the differences is limited for turnover levels, since the median value of the *sad* is 1.07% for all the matched observations (including those for which the turnover values from *Invind* and *Cerved* are identical) and the indicator is negatively correlated with the number of employees. For the average number of employees the indicator is higher, with its median value equal to 3.08% for all the matched observations, but it still decreases with firm size. Since we will use only a categorization of this variable in our developments, the risk generated from having different employment values in *Invind* and *Cerved* for the same unit is contained.

A remaining concern is that the signs of the differences might follow systematic patterns. Leaving aside the cases where $inv_{j,i,t} = cer_{j,i,t}$, we probe into the issue by looking at the successions of cases for which $inv_{j,i,t} > cer_{j,i,t}$, or $inv_{j,i,t} < cer_{j,i,t}$, in order to verify whether they are random or follow systematic patterns. First we create a binary variable:

$$d_{j,i,t} = \begin{cases} 0, & \text{if } inv_{j,i,t} < cer_{j,i,t} \\ 1, & \text{if } inv_{j,i,t} > cer_{j,i,t} \end{cases}$$

We then randomly sort the observations and finally we measure the nonparametric Wald-Wolfowitz test (Hollander and Wolfe 1973) separately for the average number of employees and the turnover each year from 2002 to 2013. The high sample size allows us to use the asymptotic normal distribution to compute the test. With a significance level of 1%, we cannot reject the null hypothesis of randomness of the successions of 0 and 1 for

Table 4.   Standardized absolute difference (sad) between Invind and Cerved figures for the matched observations (2002–2013).

Average number of employees				
	Percentiles			
	25%	50%	75%	90%
Class size				
20–49	2.30	7.69	210.26	216.00
50–99	1.05	2.99	9.27	88.66
100–199	0.00	1.72	5.94	20.62
200–499	0.00	0.98	3.83	13.17
500–999	0.00	0.70	3.33	11.73
1,000 and over	0.00	0.57	3.29	14.80
Total	**0.22**	**3.08**	**13.08**	**210.00**
Number of matched observations	42,940			
% of matched observations with $sad = 0$	23.8			

Turnover				
	Percentiles			
	25%	50%	75%	90%
Class size				
20–49	0.18	1.24	4.78	14.20
50–99	0.18	1.12	4.20	12.70
100–199	0.10	0.95	3.90	11.45
200–499	0.06	0.81	3.56	11.55
500–999	0.03	0.76	3.89	12.90
1,000 and over	0.02	1.06	5.32	15.76
Total	**0.13**	**1.07**	**4.27**	**13.07**
Number of matched observations	42,846			
% of matched observations with $sad = 0$	11.0			

Source: Bank of Italy's Invind business survey and Cerved archive.

$d_{j,i,t}$ respectively in nine and eleven of all the 12 years considered for the turnover and for the average number of employees (results of the test are available on request).

We therefore conclude that the quality of the matching is acceptable.

5.2.3.   The Imputation of the Employment Levels

A necessary step is the imputation of the average number of employees for the firms of the *Cerved* archive without this figure. We need the workforce levels for all the firms belonging to the *Invind* reference population (formed by the firms with at least 20 employees operating in the economic sectors of interest) in order to model the probability to enter the cross-sectional sample, since a categorization of this figure is a control variable in the model.

We use the predicted values of a linear model estimated by Ordinary Least Squares (OLS) to impute the missing data. We estimate the model only for the firms with turnover greater than a threshold represented by the lowest turnover found in the survey minus 20%.

This restriction has two aims: 1) preventing the model from being excessively influenced by smaller firms; 2) using units with similar probabilities of having missing number of employees, since firms below the threshold tend to be more affected by the problem. The dependent variable in the model is the average workforce level for the year, the covariates include some economic indicators (ROA, ROE, turnover), together with economic sector of activity and geographical location of firm's head office.

We select the best model within a set of possible specifications according to the following criteria:

- the covariates should not have a high frequency of missing values;
- the specification should be reasonable according to basic economic theory, even though no causal modeling is attempted;
- the fit to the data should be good (in terms of *R-square*);
- estimated totals of firms and employees for the most relevant classification cells should be close to those of the official aggregate evidence available from the Statistical Archive of Active Enterprises (ASIA), provided by the Italian National Statistical Institute (Istat 2014).

We estimate the selected model separately within cells (Table S2 in Supplementary material) formed by the combination of 20 analytical economic activities (instead of eleven ones used in the stratification) and geographical locations of the firm's headquarters (with the same level of detail as in the post-stratification). Using separate models with the same specification substantially improves the values of the adjusted R-square (it ranges between 0.7 and 0.9 for the various cells), with respect to the alternative of using the economic activity and the geographical location as covariates in a unique regression, with common parameters for the quantitative regressors. We report the model equation below:

$$Employees_{t,j,i} = \beta_{0t,j} + \beta_{1t,j}turn_{t,j,i} + \beta_{2t,j}turn^2_{t,j,i} + \beta_{3t,j}roa_{t,j,i} + \beta_{4t,j}roe_{t,j,i} + \varepsilon_{t,j,i} \quad (3)$$

where t, j, i and *turn* respectively indicate year, cell, individual firm and turnover. According to economic theory, labor cost is a better predictor than turnover for the number of employees in specifications like those shown in Equation (3). Unfortunately, we cannot use them because they are missing in 29% of cases. Firms' turnover is however a reasonable proxy, because it is highly correlated with labor costs within the cells where we estimate the model (the correlation coefficient is 0.85 on average). By using this estimate to impute the missing values for the number of employees, we obtain distributions of enterprises and employees (for the firms with at least 20 employees) similar to those of the official sources for this population (Table S3 in Supplementary material).

We use multiple imputation to estimate the variability of the imputation process by creating ten independent replications of the model predictions, each obtained by adding to the model prediction a random drawing from the residual distribution.

6. Enterprises' Performances and Panel Attrition in the Invind Survey

6.1. First Evidence of the Attrition Effects

For every survey edition, we calculate the simple and aggregate indexes presented in Table 2, weighted with the survey weights, for turnover and profit changes, ROA and

ROE, over three groups: new entrants (absent in the previous edition), future dropouts (absent in the next edition) and the rest of the sample. We express them as percent relative changes by subtracting one and multiplying the result by one hundred.

The effect of panel attrition is clearly visible for the simple indexes that are greater for new entrants than for future dropouts (Figure 2). This is true on average over the whole period, as well as for all the single years, with the exception of 2006, 2010, and 2012 for profit changes. For the aggregate index, the distinction between new entrants and future dropouts is less clear-cut, since the indexes relative to the latter ones are higher than those for the former ones for four years of the turnover changes series, five years of the ROA series and again for four years of the series of profit changes. Only for ROE is the index for new entrants always greater than that for future dropouts, even if, on average, the two are slightly closer than in the case of the simple index (the average distance is 4.8 points against 5.0).

It is now relevant to determine whether the differences discussed above are statistically significant. With this aim, for every survey edition, we separately regress each of the four indicators considered on a binary dummy identifying the two groups of interest, controlling at the same time for the survey design variables.

For every year t, we estimate the following linear model:

$$y_{t,i} = \beta_{0t} + \beta_{1t}d_{t,i} + \underline{\beta}_t\underline{X}_{t,i} + \varepsilon_{t,i} \tag{4}$$

where $y_{t,i}$ is one of the four indicators considered, $d_{t,i}$ is the binary dummy indicating whether the unit is a new entrant or a future dropout and $\underline{X}_{t,i}$ is an array of firm-level characteristics (firm size, economic sector of activity, headquarters' geographical location). We also estimate a synthetic version of the model (4) on the pooled data set and we finally replicate the regressions separately on the firms with 20–99 employees and on those with 100 employees and over.

In the year-by year regressions over the firms with 20 employees and over, the coefficients relative to the dummy are almost always positive and significant for turnover changes, ROE and ROA (Table 5). For profit changes, the effect of the dummy is always positive, but is statistically significant only for one year. The dummy coefficient is always positive and highly significant also in the corresponding regressions over the pooled data. The regressions on the two separate groups of firms with 20–99 employees and 100 employees and over reveal that, for the group of bigger firms, the differences between new entrants and future dropouts attenuate and the shares of new entrants and future dropouts decrease.

These results are consistent with the descriptive evidence of Figure 2, since they show that the economic performances of the firms not regularly participating in the survey tend to be more similar when firm size increases, even after conditioning with other observable characteristics.

6.2. Modeling Panel Attrition

Modeling the propensities to enter a given survey wave and to drop out of it helps us explain what drives this behavior. For this aim, we estimate a *logit* model over a pooling of all the observations. We use the waves for the years from 2003 to 2013 to analyze the new entrants, those from 2002 to 2012 for the dropouts. ROE, ROA, turnover and profit changes are our main covariates, to which we add dummies for the years, to take into
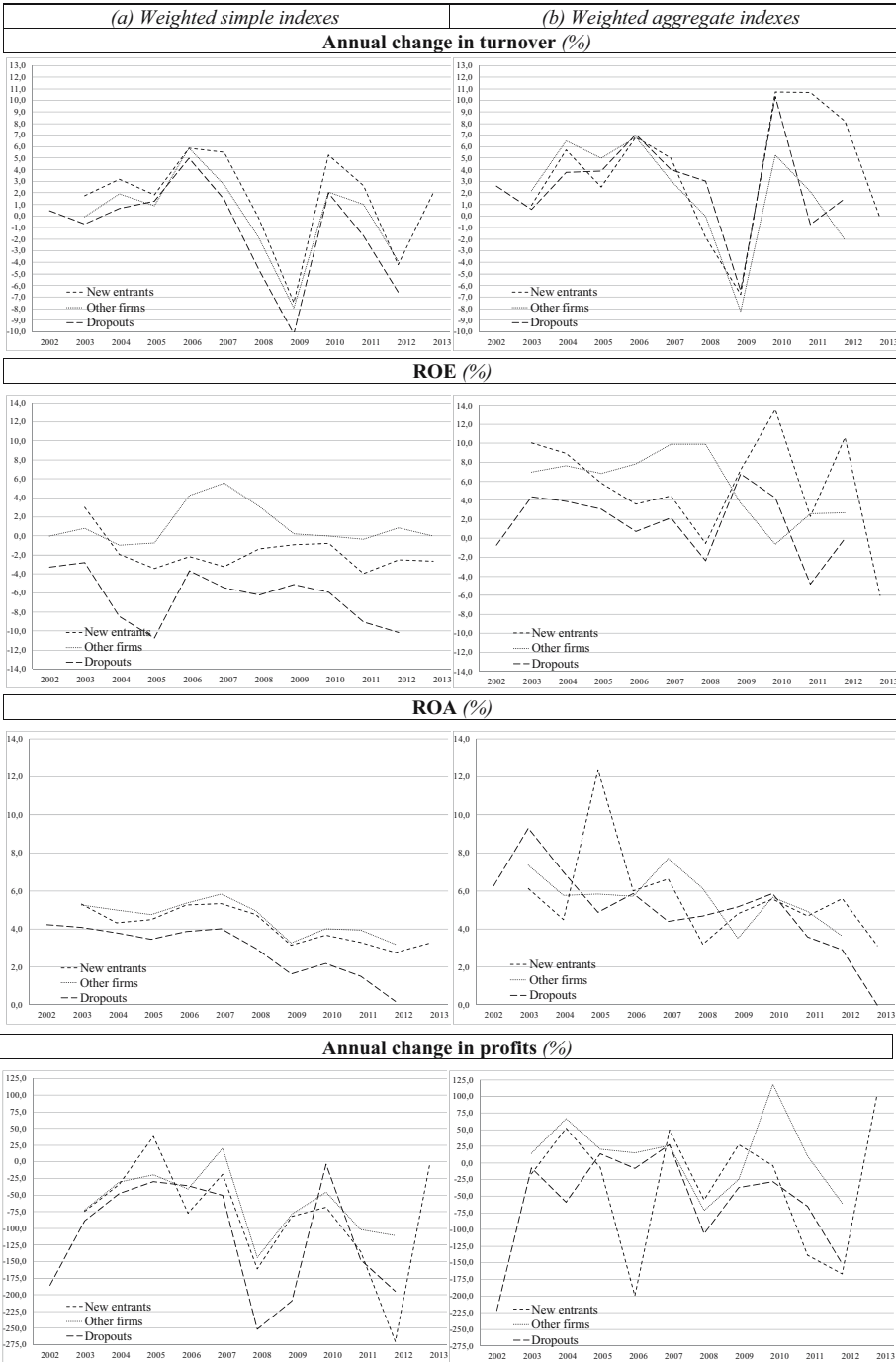
Fig. 2. Weighted indexes (percent relative changes) for new entrants in each survey edition, dropouts in the following edition and rest of the sample. Extreme values for individual values trimmed at the 1st and 99th percentiles. Firms affected by mergers, acquisitions and splits not considered. Source: Bank of Italy's Invind business survey and Cerved archive.

Table 5. Invind survey- OLS with the economic indicators as dependent variables Coefficients of the dummy for new entrants against next dropouts (2003–2012)[a][b][c][d].

Units considered: new entrants and future dropouts in each survey wave

| | Firms with 20 employee and over | | | | Only firms with 20–99 employees | | | | Only firms with 100 employees and over | | | |
| | Dependent variables | | | | Dependent variables | | | | Dependent variables | | | |
	Change in turnover	ROE	ROA	Change in profit	Change in turnover	ROE	ROA	Change in profit	Change in turnover	ROE	ROA	Change in profit
2003	4.754***	11.646***	1.756***	11.717	5.602***	13.787***	1.465**	2.260	3.370**	7.612	2.486***	17.304
2004	3.516***	8.539***	0.698	8.964	3.523**	11.721***	1.010	18.118	3.714**	1.592	−0.161	−10.509
2005	5.032***	7.253***	1.144***	49.654	5.446***	7.939***	1.213***	134.435*	3.889***	6.196	0.969	−132.132
2006	2.842***	3.144	1.578***	−89.084	4.028***	3.304	1.665***	−75.658	−0.107	3.008	1.187	−126.148
2007	4.273***	8.320***	1.639***	17.152	5.400***	5.942	1.324***	52.456	2.147	12.851**	2.334***	−38.408
2008	4.432***	4.206*	2.097***	91.924	4.543***	5.059	1.962***	157.863	4.491	2.752	2.543***	−0.776
2009	2.857**	8.326***	1.916***	169.874**	3.401**	9.018***	1.887***	135.254	1.117	6.858	1.802*	218.863*
2010	5.806***	3.040	1.364***	−7.883	7.297***	4.930	1.318**	16.961	2.977	−0.500	1.613*	−60.449
2011	4.916***	9.376***	1.872***	106.387	7.004***	8.271**	2.003***	29.780	1.467	12.133***	1.763**	271.930**
2012	4.041***	13.570***	3.807***	29.590	4.000**	14.757***	4.103***	97.645	4.355**	11.148**	3.159***	−127.287
Pooled estimates[e]	**4.280***	**6.211***	**1.859***	**38.830**	**4.900***	**7.036***	**2.007***	**62.800***	**3.200***	**4.979***	**1.624***	**0.010**

Average share of new entrants and future dropouts over the total sample

	Firms with 20 employee and over	Only firms with 20–99 employees	Only firms with 100 employees and over
New entrants (A)	**17.1**	**21.9**	**18.0**
Future dropouts (B)	**9.2**	**13.9**	**11.9**
(A) + (B)	**26.3**	**35.8**	**29.9**

(a) Dependent variables are expressed in percent. The table shows the coefficients and the significance of the dummy for new entrants into the sample, with future dropouts from the sample as reference. – (b) Separate OLS estimates. – (c) ***: p-value less than 0.01, **: p-value between 0.01 and 0.05, *: p-value between 0.05 and 0.1. Standard errors estimated with White's correction for heteroskedasticity. – (d) Additional covariates: sector of economic activities, location of firm's head office, class size. – (e) Standard errors are computed by considering the same firm repeated over time as a cluster. Additional covariates: dummies for the survey years.
Source: Bank of Italy's Invind business survey and Cerved archive.

account the time effect, as well as dummies for the survey design variables, that also control for the possible nonignorability of the survey design.

6.2.1.   Modeling the Propensity to Enter a Wave

We estimate the probability to enter a wave relative to the year t. The firms modeled are those with at least 20 employees in the year $t - 1$ in the *Cerved* archive, provided they did not participate in the wave $t - 1$. Our dependent is a binary variable, indicating whether a firm will be in the wave relative to year t. The economic indicators used as covariates (we also use per capita turnover and profits) are modeled one at a time. Since each indicator is relative to the year $t - 1$, before an eligible unit enters the sample or still remains outside, it contributes to causally explain the propensity to enter the wave.

The standard error of the coefficients accounts both for the repeated observations relative to the same firm in the pooled data set and for the variability generated by using imputed values for the missing number of employees. The number of employees is required because it is a model covariate and also because it is needed to identify whether a firm belongs to the reference population of firms with 20 employees and over. The standard error of each coefficient β_k can therefore be written as:

$$SE(\beta_k) = \sqrt{V_M(\beta_k) + V_I(\beta_k)} \tag{5}$$

where $V_M(\beta_k)$ is the usual variance of the *logit* model coefficients, after correcting for the clustering effect derived by repeatedly using the same firm (Rogers, 1993) and $V_I(\beta_k)$ is the variance of the coefficient estimates computed over ten independent replicates of the imputation process.

The coefficients for ROA, ROE, profits change and per capita profits are positive and significant (Table 6). For turnover, the evidence is mixed, since the coefficient for per capita turnover is positive and not significant, whereas the one relative to turnover change is negative but weakly significant.

The overall result is that favorable economic indicators tend to positively associate with the propensity to enter a wave, with the strongest significance measured for ROA and ROE. The sign and the significance of the coefficients relative to the dummies for firm size (Table S4 in Supplementary material) indicate that, for the firms with less than 200 employees, the propensity to enter the sample is lower than for the firms with at least 500 employees. A geographical effect also emerges, because new entrances are more easily found among firms with head office in the two macro-areas of Italy - Center and Italy - South and isles, compared with those headquartered in the macro-area of Italy - North. This effect is also due to the sampling rate for firms headquartered in the macro-area of Italy - North structurally lower than that of the rest of the sample.

A relevant question is whether the performance gap between new entrants and rest of the reference population persists, once the latter ones enter the sample. We evaluate the issue by estimating the same *logit*, with the only difference that the economic indicators refer to the year when the new entrants enter the sample, instead of the previous one (Table 7). The positive relationship between the propensity to enter a wave and positive economic performances weakens, since the sizes of the positive coefficients for ROE and ROA decrease; those for profit change and per capita profit also decrease to the point of

Table 6. Invind survey: logit for the propensity of firm's entrance into the next wave (2002–2012)[a].

	(1)	(2)	(3)	(4)	(5)	(6)
Economic indicators						
roa	0.01060***					
roe		0.00185***				
change in profit per employees (%)			0.00274**			
change in profit (%)				0.00305*		
change in turnover per employee (%)					0.00134	
change in turnover (%)						−0.03680*
Number of observations	954,792	924,484	538,836	669,012	562,645	693,179

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.

(a)***: p-value less than 0.01, **: p-value between 0.01 and 0.05, * = p-value between than 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster and by accounting for the variability of the model used to impute the number of employees when missing. All estimates include dummies for the years.

Source: Bank of Italy's Invind business survey and Cerved archive.

Table 7. Invind survey: logit for the relationship between firm's entrance into a wave and firm's characteristics (2003–2013)[a].

	(1)	(2)	(3)	(4)	(5)	(6)
Economic indicators						
roa	0.00729***					
roe		0.00075***				
change in profit per employees (%)			0.00014			
change in profit (%)				0.00102		
change in turnover per employee (%)					−0.27500***	
change in turnover (%)						−0.11980***
Number of observations	845,013	784,121	578,360	821,253	604,420	849,515

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.
(a)***: p-value less than 0.01, **: p-value between 0.01 and 0.05, * = p-value between than 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster and by accounting for the variability of the model used to impute the number of employees when missing. All estimates include dummies for the years.
Source: Bank of Italy's Invind business survey and Cerved archive.

becoming not significant; as for turnover, the coefficient relative to the change goes negative, whereas the one of per-employee turnover remains negative and increases in size (Table S5 in Supplementary material reports the complete details).

How to explain this result? The quick search for firms that replace those unwilling to further participate is a common need for all panel business surveys, since the interviewers face strict time constraints. *Invind* is not an exception. The firms with above-average economic performances in the most recent years are naturally easier to enrol, since their managers can devote more time to fill the complex survey questionnaire, but they later tend to be more similar to comparable firms, as sometimes measured in the economic literature of firm behavior (Knapp et al. 2006).

6.2.2. Modeling the Propensity to Drop Out of a Wave

In order to model the propensity to drop out of a wave, we consider all its units. Following the same steps as in the propensity to be a new entrant, the dependent for the *logit* is a binary variable indicating whether a firm in wave t also participates in the following wave $t + 1$. The economic indicators used as covariates are relative to the year t, so that they can explain firm behavior at time $t + 1$. In the analysis, the number of employees is never missing, since it is collected in the survey, therefore the standard errors of the coefficients must only be corrected for the repeated observations of the same firm, derived from using the pooled data set.

The results (Table 8) are a mirror of those relative to the propensity to enter the sample: negative economic performances in a given year augment the propensity to leave the sample in the following year. The coefficients with the greatest sizes are those of ROA and ROE. Survey participation is more erratic for the firms with less than 200 employees (the effect is even greater for those with less than 50 employees, as seen in Table S6 in Supplementary material). A geographical effect is also present, since the enterprises headquartered in the the macro-area of Italy - Northeast tend to exit the survey more frequently than those belonging to the reference macro-area of Italy - South and isles. The greater size of the subsample of the firms headquartered in the Northeast, compared to those of the other areas, helps contain the risk of excessively reducing the sample size for this area.

7. The Effects of the Delays in Updating the Reference Population

The population distribution is required to compute the survey weights, which are necessary for obtaining sample estimates representative for the reference population. The official distribution of Italian enterprises is available with a two-year lag. This means that the population of the year $t - 2$ is used for the first estimates, released at the end of May of the year $t + 1$, relating to the survey for year t, carried out in the first months of $t + 1$. In the following years, the survey weights are re-computed as soon as the updated population distributions become available.

The delay can have relevant effects on the estimates because of their structure. As shown in Section 3, the survey relating to year t collects the values for the main interest variables for the years $t - 1$, t, and $t + 1$, so that relative changes ($t/t - 1$) and ($t + 1/t$) are computed by using data only from this wave. The procedure would require three distinct reference populations for the three years, but the approach is not followed, since

Table 8. *Invind survey: logit for the propensity of firm's exit from the next wave (2002−2012)[a].*

	(1)	(2)	(3)	(4)	(5)	(6)
Roa	−0.02440***					
roe		−0.00390***				
change in profit per employees (%)			−0.00501***			
change in profit (%)				−0.00504***		
change in turnover per employee (%)					−0.24800***	
change in turnover (%)						−0.44670***
Number of observations	38,708	37,732	34,828	36,027	41,275	41,279

Additional controls (coefficients not included in the table): class size, headquarters' location, branch of acivity.

(a) ***: p-value less than 0.01, **: p-value between 0.01 and 0.05, *: = p-value between 0.05 and 0.1. Regressors values less than 1st or more than 99th percentile are set equal to reference percentile. Standard errors are computed by considering the same firm repeated over time as a cluster. All estimates include dummies at level of year.

Source: Bank of Italy's Invind business survey and Cerved archive.

the survey weights would have to be repeatedly updated over the years. For practical reasons, only one set of survey weights is used for each survey edition, relative to the year *t*, which are updated as soon as the correct population information becomes available, with a very limited change in the original estimates.

Looking at the trend in the survey reference population, its size remained stable until 2006 (Figure 3), but deep demographic changes took place in the following years, since the population size increased in 2007–2008. A slow and steady decrease subsequently occurred during the years 2009–2012 of the economic crisis. The downsizing of the nonconstruction industrial sector in the years 2001–2012 was partially offset by the growing number of nonfinancial service-sector firms. This fact is largely due to the structural changes of the Italian economy, which led to a decrease in the share of GDP produced by the industrial sector. To a lesser degree, it also led to the changes of the classification criteria of economic activities, which shifted a range of activity (such as product maintenance and customer support) previously classified as industrial to the services sectors (Istat 2010).

Given these population changes, it is important to measure the error caused by using a unique set of weights. Our analysis focuses on turnover, employment levels and investment collected in the survey and most widely used. First of all, we assess the aggregate indexes of change, since they are extensively commented in the form of percent variation (Bank of Italy 2017, 66).

We express the indexes as percent changes. If we use a single weighting system, the weighted aggregate index can be written as: $\left(\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t}} - 1 \right) 100$, if two separate weighting sets for the two periods $t-1$ and t are available, it becomes: $\left(\frac{\sum_{i=1}^{n_t} y_{i,t} w_{i,t}}{\sum_{i=1}^{n_t} y_{i,t-1} w_{i,t-1}} - 1 \right) 100$.
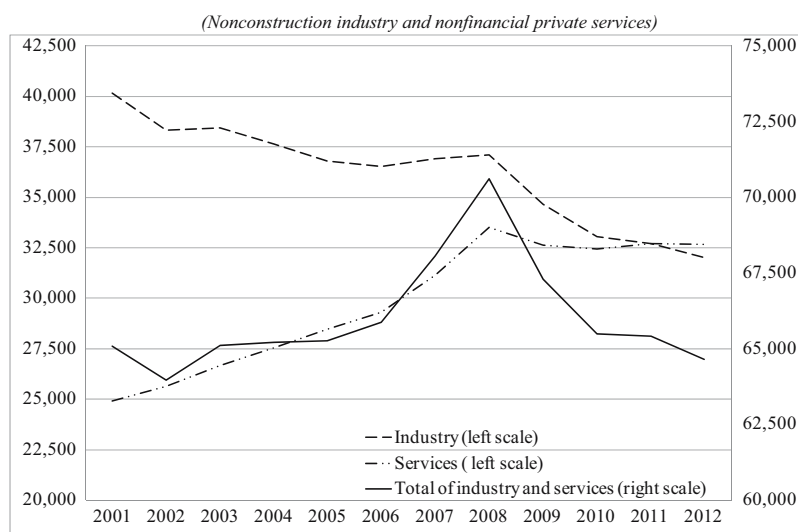

Fig. 3.   *Number of Italian firms with 20 employees and over, 2001–2012. Source: Istat, Italian Statistical Business Register (ASIA).*

It is quite easy to verify that the weighted aggregate index with the single set of weights is smaller than the one with two sets of weights whenever the population size increases, whereas it is bigger if the population becomes smaller (see last page of Supplementary material for details).

Looking at the population trends (Figure 3) and their effects on these estimates (Figure 4, upper panel), the downward bias caused by the single weighting system was strong for the years 2007–2009 (especially for investment change in 2008), when the biggest population increase occurred. In the following years, the upward bias was small, because the decrease rate of the population size was steady, but rather limited. The greatest bias shows up for the relative change of employment, as expected, since it is the estimate most affected by repeated updates of the population distribution.

We finally deal with the effect of a unique weighting system on the simple indexes, since applied econometricians use individual changes derived from the survey in their microeconometric models. These models can also be estimated in a weighted version (Cameron and Trivedi 2005, 817), in order to produce estimates valid for the whole reference population. In such a case, it is relevant to assess the effect produced on the average of these individual changes by a weighting system not adequately representative of such a population. We focus our attention on the simple indexes of the changes of turnover and employment and we disregard investment changes, which are very erratic and unsuitable for microeconometric modeling (Doms and Dunne 1998).

We again express the indexes as percent changes. If we use a single weighting system, the weighted simple index can be written as: $\left(\frac{1}{\sum_{i=1}^{n_t} w_{i,t}} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} w_{i,t} - 1 \right) 100$. With two distinct weighting sets for $t-1$ and t, it can be expressed as:

$$\left(\frac{1}{\sum_{i=1}^{n_t} w_{i,t-1 \cap t}} \sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} w_{i,t-1 \cap t)} - 1 \right) 100 \tag{6}$$

Here, $w_{i,t-1 \cap t}$ is the weight referred to the firms belonging to the reference population both at time $t-1$ and at time t. We compute these weights by considering the population size $N_{h,t-1 \cap t}$ of every stratum h as the minimum between those of the periods:

$$N_{h,t-1 \cap t} = min\{N_{i,t-1}, N_{i,t}\} \tag{7}$$

This expression is an upper bound of the true value, with a negligible approximation error if the number of entrances and exits in the population are small relative to the population size in the two periods. For the reference population of *Invind*, the yearly balances of entrances and exits over the population are worth, on average, 4% for the strata and 5% for the poststrata considered in the survey design (the corresponding median values are 1.1% and 1.7%).

For turnover and employment changes, the double weighting system has a negligible effect on the simple indexes (Figure 4, lower panel).

An explanation of this result can be found by writing the difference between the same index computed respectively with the single and the double weighting system. For the
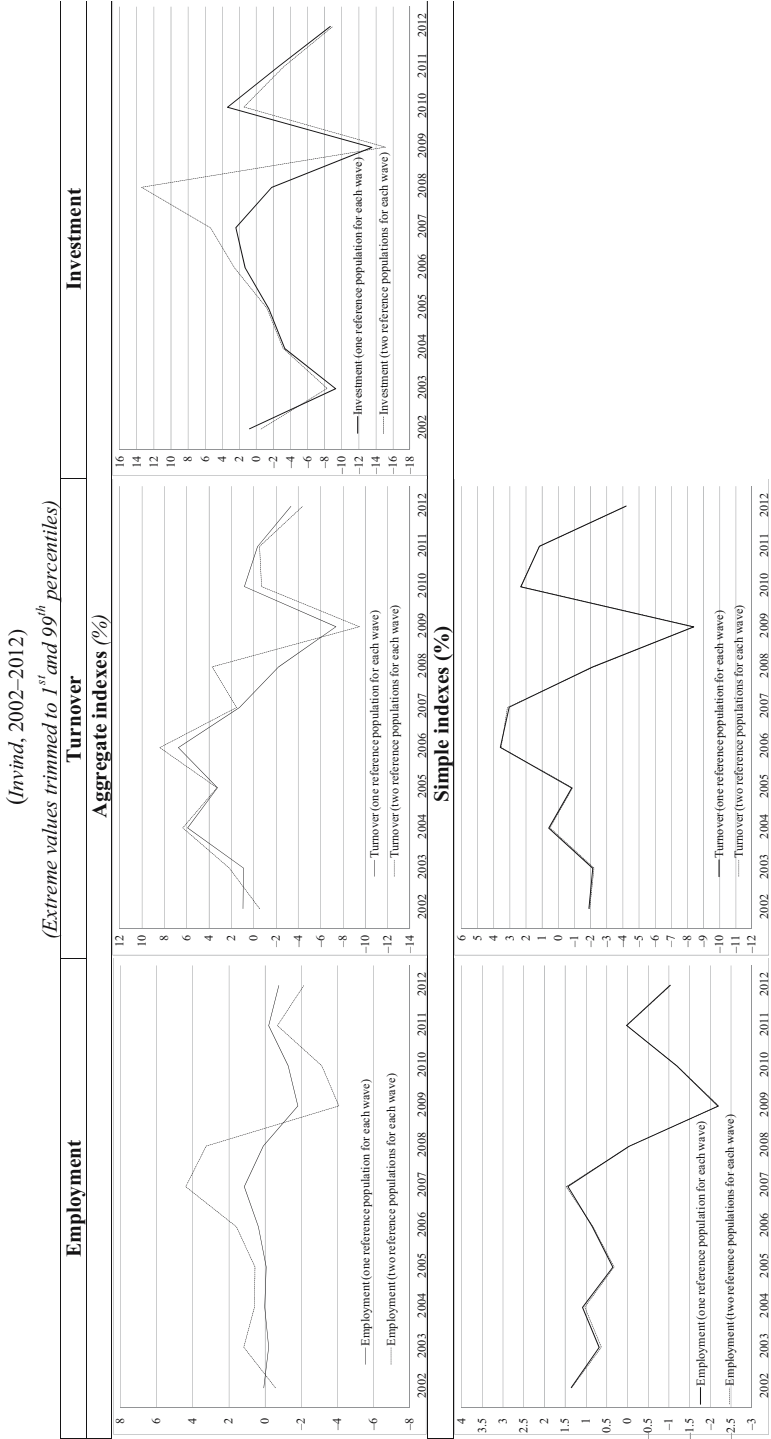
Fig. 4.   Weighted indexes (percent relative changes) of the annual relative change of employment, turnover and investment Indexes expressed as per cent changes. Turnover and investment measured at constant 2013 prices. Source: Bank of Italy's Invind business survey.

weighted simple index, this expression is:

$$\sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \left(\frac{w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t}} - \frac{w_{i,t-1 \cap t}}{\sum_{i=1}^{n_t} w_{i,t-1 \cap t}} \right) \tag{8}$$

It is a sum of n_t terms, each formed by the product of two parts: the first one $\frac{y_{i,t}}{y_{i,t-1}}$ is always positive, the second part is small, since it is the difference between the ratios of the shares of single weights over corresponding weight totals, which tends to produce very low values for Equation 8.

Looking at the weighted aggregate index, the difference becomes:

$$\sum_{i=1}^{n_t} \frac{y_{i,t}}{y_{i,t-1}} \left(\frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t} y_{i,t-1}} - \frac{y_{i,t-1} w_{i,t}}{\sum_{i=1}^{n_t} w_{i,t-1} y_{i,t-1}} \right) \tag{9}$$

The structures of (8) and (9) are similar: they sum n_t terms, each formed by the product of two parts. The first one is $\frac{y_{i,t}}{y_{i,t-1}}$, always positive, the same in both expressions. The second part within round brackets is not necessarily limited to small values in (9) and the size of the expression accordingly tends to be greater than that of Equation 8.

8. Conclusions

Our article evaluates the sampling bias of a panel business survey by using auxiliary information derived from administrative data. We consider the Bank of Italy's *Invind* survey, for which two external sources are available:

- an archive of firms' financial statements, providing complete information for all the years 2002–2013 considered for every unit participating in the survey, regardless of the continuity of its participation;
- aggregate data on the sample reference population, available for the years between 2002 and 2012.

We use the first source to evaluate the bias caused by panel attrition on the indexes of yearly changes of variables such as turnover and profits and on averages of composite indicators like ROA and ROE. We focus on the attrition caused by firms leaving the panel in a given survey edition, replaced by others in the following one.

We find the estimates that are strongly dependent on big firms' values are less affected by panel attrition than the estimates representing the average behavior of firms that do not take firm size into account.

Looking more closely at the attrition determinants, positive economic performances make it easier to enroll new firms in the survey, in order to replace firms dropping out because of negative economic performances. However, the economic results of new entrances become more aligned to those of the population, once they enter the sample.

The statistical literature (Deville et al. 1993; Särndal and Lundström 2005) proposes many adjustments to the survey weights, some of them specifically designed for business surveys (Lavallée and Labelle-Blanchet 2013). In the form of generalized post-stratification procedures, these adjustments could be useful in compensating for the bias caused by panel attrition (see also Faiella 2010 and Solon et al. 2015, for a comprehensive

review of the cases when the utilization of survey weights is advisable). For example, these techniques could be studied in future research to produce new survey weights that incorporate the information on sample firms' past economic performances.

Our analysis also suggests some possible revisions of the survey management First of all, more efforts could be devoted to prevent firms with negative economic results from dropping out of the sample at the current rate. With this aim, better interviewer training could highlight the importance of maintaining the most difficult units in the sample.

We finally analyze the effects of using survey weigths computed with a population distribution that is not up-to-date. The issue is relevant, since a single *Invind* edition collects data on the values of the interest variables for three consecutive years $t - 1$, t and $t + 1$ ($t + 1$ is a forecast that refers to the year when the interviews are conducted). The current solution is to use only the information relative to the reference population at time t, which is subsequently updated. We show how this solution entails the risk of bias for the weighted aggregate indexes of relative changes of turnover, investment and employment levels. This bias tends to be negative when the population size increases, but it becomes positive when the population size decreases, without substantial effects on the simple indexes, which do not depend on firm size. This risk can be avoided by computing, for every survey editions relative to the year t, separate sets of weights for the different years considered. This is feasible for older survey editions, while, for the most recent ones, external evidence on the population trends should be carefully used to assess the quality of the estimates.

9. References

Afonso, L.M. 2015. "Correcting for Attrition in Panel Data Using Inverse Probability Weighting: An application To the EU15 Bank System." *Doctoral dissertation (Lisbon School of Economics and Management, Working Paper)*. Available at: https://www.repository.utl.pt/bitstream/10400.5/8155/1/DM-LMA-2015.pdf (accessed January 2019).

Ardilly, P. and P. Lavallée. 2007. "Weighting in Rotating Samples: The SILC survey in France." *Survey Methodology* 33(2): 131–137.

Bank of Italy. 2005. *Supplements to the Statistical Bulletin, Sample Surveys, Survey of Industrial and Service firms*, Year 2003, Volume XV, 20 October 2005. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2003-indagini-imprese/en_suppl_55_05.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2014. *Supplements to the Statistical Bulletin, Sample Surveys, Survey of Industrial and Service firms*, Year 2013, New Series, Year XXIV, 24 July 2014. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2013-indagine-imprese/en_suppl_40_2014.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2015. *Supplements to the Statistical Bulletin, Sample Surveys Survey of Industrial and Service firms*, Year 2014, New Series, Year XXV, 1st July 2015. Available at: https://www.bancaditalia.it/pubblicazioni/indagine-imprese/2014-indagine-imprese/en_suppl_34_2015.pdf?language_id=1 (accessed January 2019).

Bank of Italy. 2017. *Annual Report for 2016*. Available at: https://www.bancaditalia.it/pubblicazioni/relazione-annuale/2016/index.html (accessed January 2019).

Black, C., D.C. Broadstock, A. Collins, and L. Hunt. 2007. "A Practical Guide to Developments in Data Imputation Methods." *Traffic Engineering and Control* 48(8): 358–363. Available at: https://www.seec.surrey.ac.uk/research/Publications/BlackBroadstockCollins&Hunt(2007).pdf (accessed january 2019).

Cameron, A.C. and P.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

Cochran, W.G. 1977. *Sampling Techniques*. New York: Wiley.

Deng, Y., D.S. Hillygus, J.P. Reiter, Y. Si, and S. Zhen. 2013. "Handling Attrition in Longitudinal Studies: The Case for Refreshment Samples." *Statistical Science* 28(2): 238–256. Doi: http://dx.doi.org/10.1214/13-STS414.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88(423): 1013–1020. Doi: http://dx.doi.org/10.1080/01621459.1993.10476369.

Doms, M. and T. Dunne. 1998. "Capital Adjustment Patterns in Manufacturing Plants." *Review of Economic Dynamics* 1(2): 409–429. Available at: http://www.homepages.ucl.ac.uk/~uctpjrt/doms%26dunne.pdf (accessed January 2019).

Fabbris, L. 1989. *L'indagine campionaria*. Nuova Italia Scientifica.

Faiella, I. 2010. "The use of survey weights in regression analysis." *Bank of Italy's Working Paper* n. 739. Doi: http://dx.doi.org/10.2139/ssrn.1601936.

Hollander, M. and D.A. Wolfe. 1973. *Nonparametric Statistical Methods*. John New York: Wiley.

Istat. 2010. Available at: http://www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007 (accessed January 2019).

Istat. 2014. *Archivio Asia*. Avaiable at: http://www.istat.it/it/archivio/archivio+asia (accessed January 2019).

Knapp, M., A. Gart, and M. Chaudhry. 2006. "The Impact of Mean Reversion of Bank Profitability on Post-merger Performance in the Banking Industry." *Journal of Banking and Finance* 30(12): 3503–3517.

Lavallée, P. and S. Labelle-Blanchet. 2013. "Indirect Sampling Applied to Skewed Populations." *Survey Methodology* 39(1): 183–215.

Little, R. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Martin, E., D. Abreu, and F. Winters. 2001. "Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation." *Journal of Official Statistics* 17(2): 267–284. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/money-and-motive-effects-of-incentives-on-panel-attrition-in-the-survey-of-income-and-program-participation.pdf (accessed January 2019).

Rogers, W. 1994. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 3(13). Available at: https://EconPapers.repec.org/RePEc:tsj:stbul-l:y:1994:v:3:i:13:sg17 (accessed January 2019).

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley.

Solon, G., S.J. Haider, and J. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* (2): 301–316. Doi: http://dx.doi.org/10.3368/jhr.50.2.301.

Trivellato, U. 1999. "Issues in the Design and Analysis of Panel Studies: a Cursory Review." *Quality & Quantity* (33): 339–352. Doi: http://dx.doi.org/10.1023/A:1004657006031.