Building panels from archives: the longitudinal representativity

Silvia Biffignandi¹ · Alessandro Zeli²

Received: 27 September 2020 / Accepted: 24 January 2021 / Published online: 5 March 2021
© The Author(s) 2021, corrected publication 2021

Abstract

In this paper, starting from ASIA business register and from the related Events Register, the longitudinally persistency of core units of the abovementioned database is examined. The M&A events are considered. To evaluate the representativity of the longitudinal panel obtained using the proposed approach, the dataset is integrated with the financial reports database with reference to some relevant economic variables. Italian financial reports database includes annual reports transmitted (according to the law) to the local Chamber of Commerce (CC) by all limited firms. The annual reports are collected and checked by CC. The database includes around 700,000 firm financial reports every year. It represents a very rich and complete source of information, therefore in this study it is used as reference target population. The complete set of data is compared both to panel based on the M&A (Mergers and Acquisitions) events and to results related to a panel build up without reference to the M&A events. Comparative analyses are carried out for the representativeness of most important economic variables. Representativeness with respect to main economic variables is analyzed. Statistical analyses, broken down by specific categories, based mainly on *R*-indicators are performed.

Keywords Research methods · Coding · Database integration · M&A events · Representativity · *R*-index

JEL Classification C60 · C88

1 Introduction

Policymakers and a range of interest groups need official statistics that allow them to make evidence-based decisions. Questions by researchers, policy makers and citizens are increasingly complex and stimulated by the dynamic relationships. In this context, there is

✉ Silvia Biffignandi
silvia.biffignandi@unibg.it; biffisil@teletu.it

Alessandro Zeli
zeli@istat.it

¹ University of Bergamo, Via Caniana, 2, 24127 Bergamo, Italy

² Istat, Viale Liegi, 13, 00184 Roma, Italy

a need to widen statistical information of official statistics in order to allow dynamic analyses and to permit the control for unobserved heterogeneity via panel econometric methods. Therefore, the availability of longitudinal data is a crucial point. In many countries, like in Italy, longitudinal data on businesses are not collected. This paper focuses on how to solve the lack of this type of data using already collected data. Reengineering the statistical business process could allow building longitudinal information while satisfying National Institute of Statistics (NSI) cost and respondent load reduction. By combining information for firms gathered in different surveys and/or from other sources a new generation of data set can be provided. The literature and some empirical studies provide experiences in using innovative approaches to create longitudinal databases. For instance, the Longitudinal Business Database (LBD) maintained by Statistics New Zealand is an important example. This database is developed by integrating various sources. This approach leads also to the opportunity of expanding information to a variety of new contents the user is interested in. See Fabling [4] for more details and for further literature.

The Research Data Center of the statistical offices of the German landers provided panel data for Germany as whole, linking surveys across time and space. Moreover, the project AFiD (Amtliche Firmendaten für Deutschland, official firm data for Germany), for details see in Malchin and Voshage [10], is merging firm level data from different surveys to build up data sets covering a wider range of variables than the ones collected in any survey, one at a time. In Italy, Oropallo and Inglese [11] and Zeli [16] proposed an approach for a longitudinal database based on annual data collection in this country.¹

In this paper, a new approach to manage the longitudinally analysed Italian core units considering the Mergers and Acquisitions (M&A) events is proposed. When an historical complex archive of data is available (for instance surveys' archive of data—or integrated archives of data—collected yearly for a long period, or administrative data sources available for a long time period as well), this mass of information may be exploited and organized longitudinally. Two possible approaches are the so-called the CIG (namely Code Identifying Group) and the COD (namely ISTAT firm's code).

Following the single firms and understanding their economic and behavioral changes is possible to create a longitudinal dataset by merely linking the firms' codes year after year (COD approach). This solution ignores the problem of M&A, which is a dramatically increasing phenomenon in the last decades (Fig. 1).

Indeed, the single enterprise status changes over time. Many events and transformation may modify the structure and other firm characteristics and its ranking on the relevant markets in terms of production and activity. Changes may affect:

- The dimensional structure (number of persons employed),
- The economic activity,
- The legal status (firm name, fiscal code).

Moreover, many events can imply the death of an enterprise. Due to possible changes the obtained longitudinal database is incorporating a greater attrition. An alternative approach consists on creating a longitudinal database that considers also the M&A (named CIG). In such a case all firms involved in M&A events are considered and they can be

¹ The general idea of this paper was presented in 2014 at the Q14 conference in Wien [1]. In that presentation only the broad outlines were supplied.

Fig. 1 Number of net cross-border M&As by region of seller, 2002–2019. Data for Italy are represented on right ordinate. Source: Unctad. World Investment Report 2015

clustered using an appropriate codification. The codification permits including in the panel longitudinal data of all relevant firms.

This paper shows the characteristic of datasets build up in Italy using these two approaches. Our hypothesis is that an improvement is gained in data quality (representativity) when the panel data taking in account the M&A events (CIG) is compared with a benchmark represented by a mere code-linked panel (COD). To verify this hypothesis, we apply different statistical methodologies to perform the comparison, such as: Spearman ranking correlation, FP-test and R-indicators.

In Sect. 2 an overview of panel data implemented in different NSI is supplied, the panels building techniques are showed in Sect. 3 the statistical methodology implemented for panels' representativity comparison is described in Sect. 4. Sections 5 and 6 present preliminary and final results respectively. Section 7 concludes.

2 Data and variables

The database used in the analysis is the result of an innovative multi-source approach. It includes the ISTAT's Business Register (ASIA), the Event Register and the Italian financial reports database.

ASIA is a complex archive. It results from the logical and physical integration of two different type of data sources, i.e. statistical sources (surveys) and administrative sources, treated with methodologies for obtaining statistical data (Tax Register, Chamber of Commerce Register, Social Security Register, Work Accident Insurance Register, Register of the Electric Power Board).

ASIA Register Service also creates the Events Register, it is a database in which all information on enterprises changes are registered.

Due the various events that can occur, a firm code may be referred to entity notably different from one year to another. The economic indicator can be strongly influenced by this.

Table 1 Main economic values of financial reports database for limited firms in Italy—years 2008–2011 (economic values in billions of euro)

Year	Firms number	Turnover	Intermedi- ate costs	Personnel costs	Value added	Persons employed
2008	687,858	2562	2587	309	512	9,220,377
2009	651,810	2241	2273	305	488	9,074,020
2010	740,516	2264	2280	308	497	8,778,157
2011	677,888	2253	2281	296	476	8,167,653

Source: Istat, Balance Sheet dataset

In order to link the firm/entity over time, the Events Register provides the following information:

- Code1, origin firm code (for instance the seller),
- Code2, destination firm code (for instance the buyer),
- Event Code, according to the above-mentioned scheme,
- Event Date, month and year,
- Quality Code.

The enterprises transformations are classified in two groups: legal status modifications and transfer of property, hence, the Event Register classification has 2 classes of nomenclatures: change of legal status nomenclature and transfer of property nomenclature. The first nomenclature includes all events of transformation of the original firm, such as: transfer by means of transformation in a new firm, cessation by means of transformation in a new firm, transfer by means of transformation in an existing firm, birth and cessation by means of a legal status change, acquisitions et cetera.

The second nomenclature includes events such as cessation by means of succession on death or takeover, transfer by means of business/enterprise lease, birth by means of succession on death or property transfer, birth by means of business/enterprise lease.

Combining the origin and destination code and considering the event code a complete registration of all possible transformation events that might occur is obtained. This information is used for panel construction purposes.

The last database, containing the Italian financial reports, collects annual reports transmitted, according to the law, to the local Chamber of Commerce (CC) by all limited firms. Due to its relevance in terms of number of observations and target population representativeness, ISTAT uses this source in many processes such as survey data validation, panel building. It is also used as backbone dataset for many simulation models. The database includes around 700,000 firm financial reports every year. It represents a rich and complete source of information.

Table 1 shows the data of the main economic variables contained in the available Istat financial reports for years from 2008 to 2011.

In particular the table presents the firms' number, the turnover (i.e., the firms' sales amount), the intermediate costs (i.e., the commodity), the personnel costs, the value added (calculated as revenues less intermediate consumption) and the total of persons employed in the firm (the sum of employees, self-employed and other people working in the firm).

3 Panels' construction

Literature and certain empirical studies encompass alternative survey approaches to collecting data longitudinally [6, 9]. Here we consider three different approaches for preparing panel data from an archival source:

- 1) A cross-sectional approach, i.e. a series of cross-sections bound together, see for instance [7].
- 2) A simple codes panel (COD panel), merely built by linking the register identification code year after year, with no regards if the economic entity remains the same;
- 3) A panel which considers firms' transformation events. A complete technical description of this approach is presented in Zeli [16].

The analysis of the effects of different approaches for a longitudinal study of the data, has to consider different factors or properties. Application criteria simplicity/easy to treat with respect to the available databases, time length of the database that can be obtained by applying the different methodology, consistency with actual facts which are registered are examples of factors/properties to be taken into account.

Cross-sectional approach is the one that fits more the real firm legal and economic development, but longitudinally linking these entities is not possible in each individual case and for every possible event. However, the cross-section approach is not considered here.

The simple code panel (COD) has the advantage of being simple in procedures and concepts, and of requiring a short preparing time. According to this approach the M&A events are not considered; therefore, there is no guarantee for the actual (both legal and economic) continuity of starting firms. Drawbacks associated with analyses and conclusions based on datasets build-up on simple code approach are important misrepresentation effects.

Longitudinalization through the first approach, i.e., clusters of firms involved in events of transformation, is considered in this paper. All M&A events are used to cluster the involved firms by means of a special code (CIG); an economic super-entity is build-up. It may be traced back to the firm(s) existing in the starting year. By means of this procedure two targets can be achieved: the economic continuity of a single firm and maximisation of the representativity of panel firms and so of the panel itself. Focusing on the transformation events approach [16], Fig. 2 illustrates the scheme followed to cluster firms with 50 employees or more in Italy. An illustrative series of events during a period of 4 years is considered to describe the longitudinal recording steps. The black codes are the new longitudinal ones created for clustering the firms involved in the M&A events, while the dark grey shadowed codes are the usual ASIA codes. At every annual transition (black, firms' clusters are created. The relevant quantitative (economic data have to be treated, for instance, summed up. For qualitative data, on the contrary, some prevalence criteria have to be adopted. In this example, three clusters are obtained; they represent the virtual firms spanning over the considered time period. Both the 6 starting firms (F1, F2, F5, F8, F10 and F12 and the 7 firms present in the last period (F3, F4, F6, F7, F9, F10 and F12 and are represented in the 3 clusters (A, B and C. If the clustering approach had not been applied and a code panel were built instead, only two firms would have been included in the code panel itself (F10 and F12 with a loss of information.

The code panel is easy to implement, but it fails to express variables levels consistent with the real ones and in granting a strong longitudinal characteristic and coverage for the

Fig. 2 Clustering events scheme (Source: reworking from [16])

firms included in the set of interest. The panel built by grouping firms involved in M&A events may be very difficult to implement starting, every time again, from the base data sources: Business Register, Events Register (i.e., Register of M&A), Business data warehouse. However, when a statistical procedure implementing this kind of panel is available (as in our case) it does not take long. As regards the strength in the longitudinal context it could be more reliable. Having as background information the comments and concepts discussed above, the aim of this paper is to evaluate the longitudinal reliability that can be achieved taking the transformation events into account while building up a longitudinal

dataset. A dataset construction by means of the grouping of firms involved in M&A events is undertaken following the scheme presented in Fig. 2. Namely the set of firms involved in one or more events is clustered by using the Event Register. Afterwards, we considered the rest of ASIA register and these firms (not involved in any events) are “clustered” by assigning a cluster number (cig) to each single firm. Subsequently, the list of enterprises obtained is linked to the economic data coming from financial reports of limited enterprises laid down to Chamber of Commerce for a time 4 years span from 2008 to 2011. Having the availability of the longitudinal dataset based on two different approaches, i.e., by grouping firms involved in transformation events and by the simply code panel approach, a comparison of these dataset in terms of representativity and coverage with respect to the whole population of financial reports (i.e., limited enterprises in Italy) is studied in this paper. From here on we denote the first panel as cig panel and the simple code panel as cod panel.

4 Methodology for the statistical comparison

In order to compare the distribution of economic variables across the databases the following analyses are undertaken.

First, Spearman’s rank correlation coefficient is applied to evaluate whether the ranking of different economic sectors remains the same in cig panel and cod panel with reference to the total population. This analysis is useful to understand if changes in the ordinal structure are present, Spearman’s rank correlation coefficients for value added, persons employed and other economic variables for the years 2008 and 2011 were calculated with reference to the data classified by economic activity.

Since Spearman’s rank correlation verifies only an ordinal ranking on the relative variations among the economic divisions, it provides no information on absolute changes. To overcome this problem, as second step of our analyses, the Fligner-Policello (F-P) test is used to verify the equality of distributions for cig panel and cod panel with respect to the universe of the relative share of the economic divisions of the economic variables considered with respect to the totals for each year. Thus, the equality of the distributions of sectorial shares of the economic variables in cig panel and cod panel with respect to total population for each year is tested using the F-P test [5, 8]. For an application in a similar context see Biffignandi et al. [3].

At the last step, an R-index is applied to evaluate the representativeness of cig panel and cod panel in terms of sector breakdown and size class. The focus is particularly on the comparison of the response composition to that of our population. R-indicators have been proposed in recent literature [12–15]. They are useful to compare the quality of different surveys or registers and apply to different situations. They can be used to compare the response to different surveys sharing the same target population to understand deviations from representativeness. The R-indicators are also useful to compare response to a survey over time, e.g., monthly, quarterly or annually. A further application is to monitor the response to a survey during data collection, in this case the representativeness of the sample during the fieldwork is checked and it is possible to adjust the data collection process.

Representativity is based on the comparison of response propensities for different units given their values on a specified set of auxiliary variables. When all response propensities in the sample are equal, the sample is representative. To compute the standardized bias the response rate is combined with the R-indicator. When comparing a survey in time, the number of sets of auxiliary variables has to be fixed in the computation of the R-indicators.

Table 2 Coverage rate of the two panel with respect to target population—years 2008–2011

	Firms	Turnover	Intermediate costs	Personnel costs	Value added	Persons employed
2008						
Cig	66.8	82.4	81.9	83.0	85.0	81.0
Cod	66.5	75.6	74.9	78.4	80.1	77.5
2009						
Cig	70.6	84.4	83.6	84.6	85.9	83.0
Cod	70.2	77.6	76.7	79.6	80.5	79.2
2010						
Cig	62.3	88.4	87.8	88.2	89.2	84.0
Cod	61.8	83.6	83.0	84.6	85.4	81.5
2011						
Cig	67.8	89.7	89.5	89.1	89.6	86.7
Cod	67.3	87.6	87.3	86.5	86.9	84.4

In this case it is possible to include also variables that correlate to the main survey items, even if they not belong to the general population characteristics. An application of this methodology in a temporal context is provided in Biffignandi and Zeli [2].

It is also possible to compute partial R-indicators. The conditional partial R-indicator measures the impact of a variable on the lack of representativity, corrected for the impact of other variables in the logit model. It is computed as the remaining within-cell variation after removal of the variable from the model. It measures ‘pure’ effects. A large value means the variable is important for explaining lack of representativity. By decomposing the variation in response propensities according to different values of the auxiliary variables, the representativeness with respect to each specific variable is obtained. The unconditional partial R-indicator for a category of a variable measures the contribution of a category to the conditional partial indicator. It measures the impact of a single variable/category on deviations from representative response. A large value means a category is a source of representativity problems. Large value can mean under- or over-representation. In all cases, partial R-indicators ideally have values equal to zero. The unconditional category-level partial R-indicators have a positive or negative sign. The sign indicates the direction of the representation of the category, e.g., a negative sign for males implies that males responded less to the survey. The conditional category level partial R-indicators do not have a sign and are always positive. In fact, in this conditional case, a category may perform better in one subpopulation but not be better in another. For example, textile sector may do better than average while food sector may do worse than average. We refer to Schouten et al. [13] for detailed descriptions of the various indicators.

5 Preliminary results

Some preliminary analyses are carried out in order to detect the possible differences in coverage rate for the event panel (cig) and the code panel (cod) with respect to the target population of whole limited firms in Italy [1]. Table 2 presents the coverage rates of the two panels with respects of target population for some important economic variables such

Table 3 Spearman correlation coefficients between total population and cig–cod panel for Nace two-digit level of some relevant economic variables—years 2008 and 2011

	2008		2011	
	cod	cig	cod	cig
Persons employed	99.6	99.6	99.6	99.6
P	<.0001	<.0001	<.0001	<.0001
Intermediate costs	95.6	96.2	99.5	99.6
P	<.0001	<.0001	<.0001	<.0001
Profits	86.9	88.3	97.8	98.6
P	<.0001	<.0001	<.0001	<.0001
Personnel costs	99.1	99.2	99.5	99.6
I	<.0001	<.0001	<.0001	<.0001
Value added	99.2	99.3	99.7	99.8
P	<.0001	<.0001	<.0001	<.0001
Turnover	95.5	96.3	99.5	99.6
P	<.0001	<.0001	<.0001	<.0001

as: turnover, intermediate costs, personnel costs, value added, persons employed and number of firms. cig represents the M&A consistent panel data and cod represents simple code linked panel data.

Every year, cig panel performs for all variables better than the cod one in term of coverage rate. The difference between the two panels with respect to firm's coverage rate is anyway rather small. From 2008 to 2011, firm's coverage rate is fairly stable in both panels. As regards the coverage rate in terms of economic variables, the difference between the two panels is a quite relevant value. The positive issue is that the coverage rate is increasing from 2008 to 2011. The critical point is that the difference in terms of rate remains high. The cig panel is better in terms of higher coverage rate. If we compute the implications in terms of value of each economic aggregate, the different performance of cig panel with respect to cod panel becomes relevant.

6 Results

This section presents the results obtained by the application of the methodology described in the Sect. 4. First of all, our approach focuses on the ranking of the sectors, which is a critical economic information to support policy decision making, and to understand the changing pattern of an economic system. Spearman correlation coefficient is providing a synthetic indicator to compare the ranking; therefore, we apply it as preliminary information for the quality evaluation of the panels. The statistical notation of this coefficient is reported in Appendix B. The Spearman correlation coefficients are calculated at the two-digit NACE level (economic divisions). For the years 2008 and 2011, the results are shown in Table 3.

Table 3 shows similar Spearman coefficients values for cig and cod panel. It means that at NACE 2-digit levels the ranking of economic sectors does not vary much from one approach to the other. Anyway, cig Spearman coefficients are higher than cod ones for any variables in both years. Therefore, implementing the grouping of firms by events improves representativity at least in terms of ranking of economic activity.

Table 4 FP-test results for the distributions of selected relevant economic variables (year 2008 and 2011)

	Turnover	Personnel costs	Intermediate costs	Value added	Persons employed
2008					
cig: FP value	0.95	0.87	0.99	0.78	1.55
p value	0.17	0.19	.0.16	0.21	0.05
cod: FP value	1.08	0.88	1.09	0.8	1.46
p value	0.13	0.19	0.13	0.21	0.07
2011					
cig: FP value	0.67	0.74	0.71	0.76	0.77
p value	0.25	0.22	0.23	0.22	0.22
cod: FP value	0.82	0.90	0.88	0.94	0.93
p value	0.20	0.18	0.19	0.17	0.27

Table 5 R-indicator for cod and cig panels—years 2008 and 2011

	R_indicator	variance_R	std_R	ci_R	LB_R	UB_R
2008 Cig	0.635	1.6E−08	0.000	0.000	0.634	0.635
2008 Cod	0.634	1.5E−08	0.000	0.000	0.634	0.634
2011 Cig	0.631	2.9E−07	0.001	0.001	0.630	0.632
2011 Cod	0.631	2.7E−07	0.001	0.001	0.630	0.632

Spearman's rank correlation verifies only an ordinal ranking on the relative changes among the economic activities. To get insights on absolute changes, as a second step of our procedure we apply the Fligner-Policello (FP) test. For the characteristics of this test references are Fligner and Policello [5] and Hollander and Wolfe [8], in Appendix B basic statistical notations are reported. FP test is useful to verify the equality of the distributions (null hypothesis) and eventually to identify which distribution is dominant (alternative hypothesis) FP tests are calculated on the two-digit NACE level (economic divisions) distributions separately for the cig and the cod panel; each panel is compared with the original population. FP-test results for the distributions for 2008 and 2011 are presented in Table 4.

FP test values and the related p values (higher than 10%) do not point out significant difference between the two panels and the original population of firms for the variables under study. Thus, the hypothesis of equality of distributions is accepted. Therefore, both the event panel and the cod panel well represent the original distribution of firms. Only "Persons employed" variable has probability slightly smaller (under 10%) for the equality of the distributions. If the values of FP test are considered as a "distance" between the panel distributions and the target population, the test value for the cig panel is in every case (but the person employed for the year 2008) smaller than the one for the cod data. It means that the cig panel is closer to original distribution with respect to the cod one. Thus, this test confirms that the cig panel approach improves representativity with respect to the cod one. To support a comparison of the quality of the two panels that share the same target population, our approach focuses on a third type of analysis: R-indicators, as a measure of representativeness. They have been proposed in literature by [12] (additional references are [13, 15], in Appendix B some statistical characteristics are summarized. The R-indexes are calculated at firm micro-data level. In Table 5 the R-indicators are reported as well as the R variance (R_variance), the R standard error (std_R), the

Fig. 3 Category level unconditional partial indicators (economic sectors)—year 2008. See Appendix A for the sectoral activity codes description

confidence interval (ci_R) and lower and upper bound for the R estimate (LB_R and UB_R). However, one has to bear in mind that R-indexes, by considering only the population of firms, display less sensitivity, with respect to economic variables, to the application of the two different panel methodologies. The results show similar values in the cod and cig panel.

To measure dissimilarity of each panel from the target population at category level of the two auxiliary variables unconditional partial indicators are computed. As regards the category level unconditional partial indicators, the two considered variables are economic sectors and size class. Economic sectors breakdown at sections of NACE rev.2 level, whereas firm size breakdown considers eight classes of number of persons employed, i.e. less than 5, 5 to 10, 11 to 20, 21 to 50, 51 to 100, 101 to 250, 251 to 1000, over 1000. Figure 3 shows the values of category level partial indicators for the economic sectors (sections of NACE rev.2) for year 2008.

Mining (B), manufacturing (C), trade (G) and other business services in L are in both type of panel over-represented while the other sectors are underrepresented. In 2008 cig panel presents higher over representativity for industrial sectors and trade with respect to cod panel. Size class in terms of persons employed partial indicators for year 2008 are presented in Fig. 4. In this case cig panel represents better the larger firms over five persons employed while it registers an underrepresentation for the smallest firms.

In year 2011 (Fig. 5) many economic sectors present an over representativity: mining (B), manufacturing (C) and water supply (E) among industrial sectors and trade (G), information and communication (J) and professional activities (M) among services.

For year 2011 the partial indicators for size class over represented all size classes but the micro-firms with less than five persons employed (the most numerous one). R partial indicators register a better representation for cod panel for the smallest firms (with less than 10 persons employed this time) and confirm for cig panel a greater over-representativity for larger enterprises with more than ten persons employed (Fig. 6).

Fig. 4 Category level unconditional partial indicators (size class)—year 2008. See Appendix A for the sectoral activity codes description

Fig. 5 Category level unconditional partial indicators (economic sectors)—year 2011

Fig. 6 Category level unconditional partial indicators (size class)—year 2011

7 Final remarks

The paper considers two alternative approaches to build up longitudinal business panels: the new multi-source approach for taking transformation events in the panel construction is proposed (cig database) and the longitudinal panel based on simple code approach (cod database). Panels' units are linked to the Chamber of Commerce balance sheet reports database. Selected relevant economic variables (turnover, personnel costs, value added, person employed) are computed with reference to each longitudinal panel and results compared.

The aim of this study is to understand if the two approaches provide estimates of the variables of valuable quality for the economic analysis and to compare their performance. A methodological approach based on a sequence of tests is proposed to answer the research question. The value added of the presented approach is to gain insights in key issues for understanding the longitudinal characteristics of the economic aggregates, namely the ranking of sectors and the distribution.

The adopted test and methodologies show that the ranking of economic sectors does not vary very much from one approach to the other, and both panels well represent the original distribution of firms. Nevertheless, looking to the results in more detail, the firms grouped by events seem to improve representativity in terms of ranking and distribution.

Considering the partial indicators for size class register, both in the 2008 and in 2011 a better representation for cod panel for the smallest firms (with less than 10 persons employed) is detected, whereas cig panel confirms a greater over-representativity for larger enterprises (with more than 10 persons employed). As regards economic sectors, partial indicators show that mining, manufacturing, trade are in both type of panel over-represented while the other sectors are underrepresented.

These results confirm that the proposed approach for panel construction taking in account transformation events performs well, even better than the simple code approach.

This study has the limitation that only few variables are processed to evaluate and compare panel quality; even if the considered variables are the most important from economic point of view, it would be interesting to check the improvement of the proposed approach over other variables. Further studies to evaluate in more detail the impact of the proposed approach to economic results are being undertaken.

Appendix A: Legenda for economic sectors' NACE codes

See Table 6.

Table 6 NACE rev.2 Section level and description

B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
P	Education
Q	Human health and social work activities
R	Arts, entertainment and recreation
S	Other service activities

Appendix B: Statistical notation of the applied methods and tests

Spearman correlation coefficient

Spearman correlation coefficient is a measure of rank correlation (statistical dependence between the rankings of two variables).

It is computed as follows:

$$r_s = 1 - \frac{6 \sum d_j^2}{n(n^2 - 1)}$$

d_i is the difference between two ranks of each observation n is the number of observations. Spearman correlation ranges between $+1$ and -1 ; value of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the others. Generally speaking, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of -1) rank between the two variables.

With respects to the comparison between two panels based on different construction procedures, high rank correlation coefficients suggest that the adopted procedure does not modify the differences in the distribution of the population variable and the estimated variables obtained using the considered procedures therefore the relative contribution of each sector is not affected by the method.

Fligner-Policello (FP) test

Fligner-Policello (FP) test is used to verify the equality of distributions. For example, FP test for the equality could be applied to the distributions of sectoral shares of the economic variables in two panels with respect to total population.

Let F_t and F_p be the distributions of the variable of interest. Let us denote with $X_t \approx F_t$ and $X_p \approx F_p$ the associated random variables and with X_t and X_p the two respective realizations. The distribution F_t is estimated to dominate F_p if $\text{Prob}[X_t > X_p] > 1/2$, that is, if one randomly

$$H_0 : \int dF_t F_p = \frac{1}{2} \quad \text{vs} \quad H_1 : \int dF_t F_p \neq \frac{1}{2}$$

this is due to the fact that

$$\text{Prob}[X_t > X_p] = \int dF_t(X) F_p(X)$$

If the null hypothesis is rejected, the sign of the FP statistics shows which of the two groups is dominant. The Fligner-Policello test assumes neither normality nor equal variances. Under this situation none of the usual tests to verify whether two groups differ in central tendency is appropriate. Furthermore, FP test does not even assume that the two distributions have a similar shape. Rejection of the null hypothesis means that the two distributions are different in probability because it is a test of stochastic equality. So, for example it is possible to statistically verify if the distribution of the share of a variable by economic sector in one panel is equal (or alternatively dominate) to the distribution of the share of a variable by economic sector in the corresponding population of firms (or in the other panel). If the null hypothesis is rejected, the sign of the FP statistic indicates which of the two distributions dominates the other.

The R-indexes for representativeness

The weak definition of representativeness is given in Schouten and Cobben [12] i.e. a response subset is representative for a categorical variable X with H categories if the average response propensity over the categories is constant, so:

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho$$

where N_h is the population size of category h , $\rho_{h,k}$ is the response propensity of unit k in class h and summation is over all units in this category.

R -index to take values on the interval $[0,1]$ with the value 1 being higher representativeness and value 0 being the maximum deviation from higher representativeness, the index can be based on standard deviation of response propensities, (R_1 index):

$$R_1(\bar{\rho}) = 1 - 2S(\bar{\rho})$$

Another solution is to use the variance of response propensity, so that the R_2 index is obtained as follows:

$$R_2(\bar{\rho}) = 1 - 4S^2(\bar{\rho})$$

where S is the measure of the standard deviation of the response probability.

The estimator for R_1 and R_2 are calculated in practice as:

$$\hat{R}_1(\bar{\rho}) = 1 - 2\sqrt{\frac{1}{N-1} \sum_1^N \frac{s_i}{\pi_i} \left(\hat{\rho}_i - \hat{\bar{\rho}} \right)^2}$$

$$\hat{R}_2(\bar{\rho}) = 1 - 4\frac{1}{N-1} \sum_1^N \frac{s_i}{\pi_i} \left(\hat{\rho}_i - \hat{\bar{\rho}} \right)^2$$

where N is the population, s_i is a selection indicator that takes value 1 if the unit is selected in the sample and 0 otherwise, π_i is the first order inclusion probability.

The response probability ρ_i is calculated by means a logit model as follows:

$$\rho_i = P(r_i = 1 | s_i = 1) = \frac{\exp(-\beta_i X_i)}{1 + \exp(-\beta_i X_i)}$$

To obtain an indicator that measure dissimilarity of single auxiliary variables conditional on the other auxiliary variables let us consider Z an auxiliary variable with H categories.

A marginal indicator [13] results from the ANOVA decomposition with respect to Z to the variance $S^2(\rho_x)$ underlying the R indicator. From the R indicator definition, the unconditional partial indicator for the variable Z is measured as:

$$P_u(Z, r_x) = S_b^2(rx | Z)$$

The larger the value of $P_u(Z, \rho_x)$ the grater the contribution of variable Z to the lack of representativeness.

The unconditional partial indicator can also be defined with reference to category h of Z :

$$P_u(Z = h, \rho_x) = \sqrt{\frac{N_h}{N}} (\bar{\rho}_{X,h} - \bar{\rho}_X)$$

The indicator $P_u(Z=h, \rho_x)$ may be positive or negative, indicating respectively over or under—representation for category h .

It is also possible to operate a normalization of the R index [12]. This operation makes the calculation of the non-response bias possible, and also permits the calculation of the maximum absolute bias of a target variable $\bar{y}$. In particular, it is possible to verify that:

$$\left| B\left(\frac{\hat{\bar{y}}}{\bar{y}}\right) \right| \leq \frac{S(\bar{\rho})S(y)}{\bar{\rho}}$$

where $\left| B\left(\hat{\bar{y}}\right) \right|$ is the bias of an arbitrary survey variable Y . It is also possible to obtain:

$$\left| B\left(\hat{\bar{y}}\right) \right| \leq \frac{S(\hat{\rho})}{2\hat{\rho}} \quad \text{and} \quad \left| B\left(\hat{\bar{y}}\right) \right| \leq \frac{1 - R(\rho)}{4\hat{\rho}}$$

Using these formulae is possible to calculate a lower bound for the R -index and an upper bound for the bias, i.e., the maximal absolute bias.

Funding Open access funding provided by Università degli Studi di Bergamo within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Biffignandi, S., Zeli, A.: Alternative panel data approaches: longitudinal representativity. Presented at Q2014 European conference on quality in official statistics, Vienna (2014). http://www.q2014.at/fileadmin/user_upload/Biffignandi_ppt_presentation_Q2014.pdf. Accessed 15 May 2020
2. Biffignandi, S., Zeli A.: Integrating databases over time: what about representativeness in longitudinal integrated and panel data? Paper presented at Q2010 statistics conference Finland, Helsinki, May 4–6 (2010)
3. Biffignandi, S., Nascia, L., Zeli, A.: Combining survey and administrative data in Italian business surveys. *J. Econ. Soc. Meas.* **41**(1), 67–83 (2016)
4. Fabling, R. A rough guide to New Zealand's Longitudinal Business Database. Paper prepared for the plenary session on access to enterprise microdata in Asia comparative analysis of enterprise data conference. Tokyo, October 2–4 (2009)
5. Fligner, M.A., Policello, G.E.: Robust rank procedures for the Behrens-Fisher problem. *J. Am. Stat. Assoc.* **76**(373), 162–168 (1981)
6. Frees, E.W.: Longitudinal and panel data: analysis and application in the social sciences. Cambridge University Press, Cambridge (2004)
7. Grazi, M., Sanzo, R., Secchi, A., Zeli, A.: The building process of a new integrated system of business micro-data 1989–2004. *J. Econ. Soc. Meas.* **38**(4), 291–324 (2013)
8. Hollander, M., Wolfe, D.A.: Nonparametric statistical methods, section 4, 2nd edn., pp. 135–138. Wiley, New York (1999)
9. Kessler, R.C., Greenberg, D.F.: Linear panel analysis: models of quantitative change. Academic Press, Cambridge (1981)
10. Malchin, A., Voshage, R.: Official firm data for Germany. *Schmollers Jahrbuch J. Context. Econ.* **129**(3), 501–513 (2009)
11. Oropallo, F., Inglesse, F.: The development of an integrated and systematized information system for economic and policy impact analysis. *Austrian J. Stat.* **33**(1–2), 211–223 (2004)
12. Schouten, B., Cobben, F.: R -indexes for the comparison of different fieldwork strategies and data collection mode. Discussion paper 07002, CBS (2007)
13. Schouten, B., Shlomo, N., Skinner, C.: Indicators for monitoring and improving representativeness of response. *J. Off. Stat.* **27**(2), 231–253 (2011)

14. Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., Skinner, C.: Evaluating comparing, monitoring, and improving representativeness of survey response through *R*-indicators and partial *R*-indicators. *Int. Stat. Rev.* **80**(3), 82–399 (2012)
15. Shlomo, N., Skinner, C., Schouten, B., Carolina, T., Morren, M.: Partial Indicators for representative response. RiSQ (representativity indicators for survey quality). Work Package 5 Deliverable 4, Version 2. RiSQ Project was financed by the 7th Framework Programme (FP7) of the European Union (2009)
16. Zeli, A.: The creation of an experimental longitudinal database of business micro-data. *J. Econ. Soc. Meas.* **43**(3–4), 143–167 (2018)