Pursuing Data Quality in the Multi-Purpose Statistical Survey for the Permanent Census of Enterprises and Adaptive Re-Contact Strategies

G. Bellini[1], M. Ballin[2], G.G. Di Paolo[1], S. Filiberti[3], A. La Rocca[2], P. Papa[3]
Istat - Italian National Statistical Institute

Abstract

The Multi-Purpose Statistical Survey for the Permanent Census of Enterprises regularly provides a detailed picture of the Italian economic system, by gathering information about enterprises' organization, innovation, digitalization, competitiveness and environmental sustainability.

Sampling design for the survey was defined in the first edition (2018): a sample of about 280K enterprises with at least 3 persons employed was selected from a population of 1M units. Analysis of results obtained helped in defining new strategies for data collection in the second run of the survey for year 2022.

Istat has been focusing on data quality in addition to overall response rate during data collection. A differentiated communication and solicitation strategy was adopted depending on the propensity of enterprises to participate to the survey. The first experience in such field was run during year 2021 with the survey "Business situation and prospects during and after the Covid-19 health emergency".

In order to guarantee a good level of accuracy for the final estimates, an interactive analysis of response rate and of the expected data quality for each domain of interest has been performed. Targeted groups of enterprises included in domains of interest for which data quality was considered "weak", were detected and appropriately treated. Main results achieved adopting such strategies are depicted.

Key words: multi-purpose survey; multi-domain and multivariate sample allocation; non-response rate; adaptive survey; targeted solicitation activity; R-indicators methodology.

1. Introduction and main objectives

The Multi-Purpose Statistical Survey is carried out in the context of Permanent Census of Enterprises. It regularly provides a detailed picture of the Italian economic system, by gathering qualitative information about several relevant topics. This survey, included in scope of economic statistics on enterprises, allows acquisition of data on several characteristics of the economic units, which are not deductible from other sources. Moreover, the survey makes it possible to get complementary statistics to those of quantitative nature that come out from other direct surveys or from administrative sources.

Through the survey e-questionnaire[4] a lot of detailed information are collected on the main dimensions

[1] Directorate for data collection - Istat - Italian National Statistical Institute (bellini@istat.it; dipaolo@istat.it). Bellini §§ 3.1; Di Paolo §§ 3.3.
[2] Directorate for methodology and statistical process design - Istat (ballin@istat.it ; allarocc@istat.it). Ballin - Della Rocca §§ 3.2 and 3.4 (jointly).
[3] Directorate for economic statistics - Istat (filibert@istat.it ; papa@istat.it). Filiberti §§ 1 (part), 4 (part), 2; Papa §§ 1 (part), 3, 4 (part).
[4] Survey e-questionnaire - the last edition of the survey was carried out in 2022-2023 - is structured in 9 sections: 1 - Ownership, Control and Management; 2 - Human Resources; 3 - Production Relationships and Supply Chains; 4 - Market;

around which revolve the strategic choices of Italian companies.

Concerning the thematic contents, the results of the survey are of great informative impact and value and are highly appreciated by the stakeholders, as it is demonstrated both on occasion of the presentation of the results of the previous edition (February 2020) and in the preparation phases of the new edition of the survey. However, it is necessary to find the right balance between information needs of researchers and policy makers, the need not to exceed the statistical burden of the contacted units, the cost constraints of the survey and the compliance with methodological rules aimed at obtaining high quality results.

However, it is undeniable that this detection tool, although it has been studied in order to limit the impact for the compiler as much as possible, it may have been demanding and time consuming for the contacted enterprises.

The data collection organizational model and strategy adopted changed over time.

Since centralization of data collection (DC) activities occurred in the National statistical office in year 2016, the Multi-Purpose Statistical Survey for the Permanent Census of Enterprises has been run twice, in year 2019 e in year 2023.

Main differences with the editions run before is that the census, that already moved from a typical census oriented approach to a survey run on sample basis (edition year 2011), revolved all his attention first on commitments with external stakeholders for the data collection phase, particularly the Chamber of Commerce, that had a relevant role in the conduction of 2011 Enterprises Census, and after investing more on professional external assistance services.

The experience acquired in the context of previous survey (2019) was helpful in many ways for the new edition of the permanent census. In fact, all activities, for instance the statistical methodologies that have been adopted, the choices made about the contents of the questionnaire, data collection, release of the results, organization as a whole, etc., have been carried out in a more efficient way, also considering the constant reduction of available resources.

Nowadays, the success of the Census, in terms of survey units participation, depends basically from the behavior of the involved enterprises and the solicitation procedures adopted. The experience done with surveys on enterprises shows that there are several factors affecting survey participation of involved units, in fact not all of them have the same behavior in terms of propensity to fill in the questionnaires. Moreover, dealing with sample survey a bias on final estimates is introduced as the non-respondent units have peculiar characteristics. That is the reason for considering total unit non-response as a component of the *non-sampling error*.

The results obtained in the previous census, showed that relatively important factors for survey participation are physical dimension and state of inclusion in the Business statistical portal (Table 1). In fact, higher response rate is recorded for units with higher number of employees, going from 57.4 percent for units with less than 10 employees to 75.4 for units with at least 20 employees. Referring to familiarity with the Portal system, units already involved in Istat surveys - and thus being already included in the portal - show higher response rate, equal to 87.4 percent, than newly included units, that only score 49.7 percent.

5 - Innovation and Digitization ; 6 - Finance; 7 – Internationalization of Production / Global Value Chain; 8 - Intelligent Specialization; 9 - Environmental and Social Sustainability.

Table 1 – Response rate per state of inclusion in the Business statistical portal and size (class of employees) – Year 2018

Size (class of employees)	New in the Portal	Old in the Portal	Total
3-10	51.3	86.9	57.4
10-20	49.0	87.2	68.3
20 and more	40.4	87.8	75.4
Total	49.7	87.4	64.0

Source: Business statistical portal for state of registration and Asia 2021 for number of employees

For this reason, a more targeted kind of communication has been adopted in the last edition of the Enterprise Census to solicit non-respondent survey units, as the final goal was not only the quantitative result (number of filled in collected questionnaire) but also the qualitative one.

2. Sampling design and estimation domains

With reference to the sample design chosen for the survey, it was decided to replicate the approach adopted in the previous edition. In fact, this methodology has proved to be particularly effective and efficient both in terms of costs incurred for carrying out the survey, and in relation to the overall statistical burden for the respondents, and with respect to the quality of the results obtained.

Since the total population of Italian companies is made up of about 4.5 million units (Business Register), in order to reduce statistical burden for microenterprises, a sample of about 280K enterprises with at least 3 persons employed - enterprises with 20 or more persons employed belong to a take-all stratum - was selected from the population of reference of about 1 million units (covering 77.6% of employment, 89.0% of turnover and 85.5% of value added).

The population of interest (unit of survey and analysis is the company as legal unit) was defined according to economic activity (companies operating in Nace Rev.2 classification: sections from "B" to "N", from "P" to "R" and divisions "S95" and "S96"), size (employment) and territory (21 administrative regions and 107 provinces are considered).

For the determination of the sample size and allocation into strata as a function of the expected sampling errors, it was decided to carry out simulations according to the following estimation domains: Macro economic activity sectors (Industry, Constructions, Commerce, Other services) by province; Nace Rev.2 Sections by Size class (3-9, 10-19, 20+ persons employed) by Region; Nace Rev.2 Divisions by Size class; Nace Rev.2 Divisions by Region; Nace Rev.2 Class.

The estimation domains planned in the survey design phase were coherent to those defined in the phase of publication of the results of the previous edition of the multipurpose qualitative survey (2019).

A simple random sampling design with a stratified stage was chosen for the survey. It allows to obtain information at a high level of detail. The strata (the stratification adopted was defined by "Nace Rev.2 4-digits * size classes of persons employed * provinces") were constructed coherently with the domains of interest.

The multivariate allocation in the strata was based on the use of auxiliary qualitative information obtained in the context of the previous multipurpose survey. At the same time, a multi-domain approach provides for the satisfaction of all the constraints imposed by the expected sampling errors (pre-set thresholds) for each of the domains of interest. The optimal size of the sample was obtained by binding on the maximum expected errors of three guiding variables measured in the context of the previous edition of the survey: 1)

Introduction of at least one innovation (product/service/process/organizational/marketing); 2) Acquisition of new resources; 3) Self-financing of the company.

The determination of the sample size was obtained as a function of the maximum sampling errors expected for the estimation of the guiding variables collected in the previous edition as a function of the domains indicated. The allocation of the sample units in the strata obtained initially was further redefined according to the response rate observed in some specific domains in the 2019 edition (Ateco division and Region) to prevent any drops in advance. Each sample unit selected from the list (reference population – Business Register) is then assigned an initial sample weight (it indicates the number of population units represented by the sampled unit). After further checks carried out with respect to the updated master data of the companies, the information letters were sent to the selected sample units.

Finally, also for this survey, in order to reduce the overall statistical burden for firms, a strategy for selecting sample units with negative coordination (with respect to the other structural surveys on firms) was adopted.

3. Data collection development and strategy

In detail, the Multi-Purpose survey carried out in the period 2022-2023 has substantial methodological and technical continuity with the 2019 edition but also with the previous 2011 edition. Nevertheless, the last edition involved an organizational review, which led to a drastic reduction in the role of the territorial network and launched some compensatory processes aimed at increasing the overall efficiency of the data collection process. The following Table 2 shows a quantification of the reduction in the role of the territorial network over time.

Table 2 - Organizational aspects of data collection in the three last editions of the Business Census

Involved structures	2011	2019	2022-2023
Chambers of Commerce	YES	NO	NO
Regional census offices	YES	NO	NO
Provincial census offices	YES	NO	NO
Field detectors (n. people)	2.257	0	0
Istat Internal Territorial Network (no. of people)	About 150	About 150	About 10
Contact center for assistance	Internal staff	External supplier	External supplier
Costs	High	Low	Very low

This trend required a set of solutions that concerned organisational, technological and methodological aspects. A first solution concerned the design of new services to support users involved in the survey. They resulted in: a) placing the survey questionnaire in the context of the Business Statistical Portal, an integrated web portal for data collection of all business surveys; b) design of a centralized inbound and outbound contact center service for professional assistance and support to the units involved in the surveys.

The data collection is nowadays done only adopting a CAWI technique and the respondents can make their requests for assistance to a Toll free Number, a unique inbound and outbound service run by an external dedicated society, either for problems encountered accessing the Statistical portal either for issue related to the questionnaire content. An Istat network of specialized personnel is also involved to provide assistance, particularly dedicated to solve issues that cannot be standardized neither solved directly by the inbound service, through the FAQ system.

A second solution concerned the optimization of the data collection processes by envisaging the application of techniques for the optimization and rationalization of the survey questionnaire and the application of

adaptive techniques aimed at reducing the distortion resulting from the missing answers already during the process of data collection. The results reported in the following paragraphs go exactly in the direction of adaptive techniques application.

3.1 Standard and adaptive solicitation activity

The standard solicitation strategy - adopted for structural survey - involves the planning of massive reminders addressed to the non-respondent units involved in the survey, that are reached through formal (certified email) and informal communications (ordinary email); moreover the final recall activity, addressed to a sub-group of most relevant or significant units, is run for a short time right before the end of the survey. Normally, one or two formal communications are sent through certified email, to which two or three informal communications are added, sent to the ordinary email of the survey units gathered from the unit's registration system in the Statistical Portal.

The timetable for the management of the solicitation plan is quite strict in principle but subject to change as solicitation strategy is linked to the trends recorded in terms of compiled questionnaire collected.

It has to be reminded that the ordinary email is available for units already recorded in the Statistical Portal, whereas for units participating for the first time to a statistical survey managed by Istat, this is information is not available at the beginning of the survey. The same happens for information on telephone number used for recall activity. This means that even if channels used to contact non-respondent units are multiple, for those approaching for the first time to the statistical system, most of the time the only information available is the one on certified email.

For the Census the approach chosen was to complement standard reminders with targeted ones, and to focus recall activity on most difficult units to be gathered.

A first case study in this field was represented by the survey called "Business situation and prospects during and after the Covid-19 health emergency" (Covid survey in the following), run in different waves in years 2020 and 2021. Since it showed to be successful with non-respondent units, a similar strategy was adopted in the Multi-Purpose survey carried out in year 2023.

In terms of calendar, during four month of data collection, Istat planned the following contact activity as massive:

- 3 different formal communication through certified email sent to all non-respondent units and a final one only to non-registered units;
- 5 different informal communication through ordinary emails;
- recall activity done on high priority unis, that were the first 5000 non-respondent units, ordered according to physical size (employees number).

Targeted reminders were sent in the central period of the data collection phase, and in two different waves, the first one started in February the 5th and the second one in March the 7th, in order to have enough time to see the effect of the reminders on response rate and to analyze them. In both waves, respondent units were analyzed and a specific list of targeted units to be solicited was extracted. All the available contact channels were utilized: certified and ordinary emails, for sending reminders, and telephone for the recall activity. As targeted units were the most reluctant to participate to statistical survey, in many cases ordinary emails and telephone numbers were not available, thus especially for telephone an extra activity of gathering

contacting information was performed either by Istat - from internally available database - and by the external society managing the outbound activity.

Referring to formal communication sent to non-registered units, another strategy adopted was to send the initial login information attached to the reminder, in order to unable units to access promptly the Business statistical portal.

The volume of units to be involved in such targeted group was mainly related to technical constraint, as the maximum number of submissions per day of certified email is around 30.000 units while it is about 20.000 units for ordinary email. Thus, 50.000 was considered a reasonable number of units to be contacted in short time by sending messages, while for recall activity 35.000 was chosen as number of units to be re-contacted. In both cases, figures refers to units included in each one of the two waves run.

3.2 Identification of targeted units in the adaptive solicitation strategy

To better understand the rationale behind the non-respondent follow-up strategy adopted for the CPUE survey, it is appropriate to recall the objectives that were considered for such phase.

The objectives of the survey here considered are the production of estimates for three types of domains (hereinafter dom_1, dom_2 and dom_3). Each of these domains determines a partition of the entire reference population.

The domains are defined as follow:

- dom_1 is obtained by crossing (Nace Rev.2 macro-sector[5], size class, province); the partition is then $D_{dom_1} = 2{,}626$ non-empty enterprise sets;
- dom_2 is obtained by crossing (Nace Rev.2 2-digit, size class, region); the partition is then $D_{dom_2} = 4{,}704$ non-empty enterprise sets;
- dom_3 is obtained by crossing (Nace Rev.2 4-digit, size class); the partition is then $D_{dom_3} = 1{,}811$ non-empty enterprise sets.

The following scheme gives a statistical description of these groups with respect to their size expressed in terms of the number of enterprises.

Scheme 1. Statistical description of the three domain types

Domain	Number of groups that constitute the partition D_{dom_i}	Minimum size of groups to be estimated	25th quantile (25% of groups contains less enterprises of these threshold)	Median (50% of groups contains less enterprises of these threshold))	Average number of enterprises by group	75th quantile (25% of groups contains more enterprises of these threshold)	Maximum size of the groups to be estimated
dom_1	2,626	1	22	66	106	142	4,623
dom_2	4,704	1	7	24	59	67	1,993
dom_3	1,811	1	28	84	154	176	4,148

A possible approach to define a reminder strategy is the one suggested by the *R-indicators*.

[5] Nace Rev.2 macro-sectors: Industry, Construction, Trade, Other Services.

Reminding that the functional relationship between the bias of the following mean estimator:

$$\hat{\bar{y}} = \frac{\sum_{i=1}^{n} w_i y_i / p_i}{N}$$

where *n* is the number of sampled units, w_i indicates the weight defined by the sampling design and with p_i the response propensities; the R-indicators methodology moves from the relationship between bias of previous estimator and covariance of variable of interest and the response propensities:

$$Bias(\hat{\bar{y}}) = \frac{Cov(y,p)}{\overline{p}}$$

The suggested strategy is to lower the variability of *p* in order to lower the maximum threshold of bias. In principle, the goal is therefore to implement all possible strategies so that all units are equally likely to respond. If you can achieve this then you eliminate any self-selection phenomenon (that create bias).

If only one type of domain were considered (for example dom_1) this approach would consider each group individually.
As can be seen from the scheme above, many of these domains have a small size and therefore the estimation of the models underlying the *R-indicators* methodology could meet some difficulties.
A more pragmatic approach was therefore chosen in which the objective of the reminder strategy was to try to obtain at least some respondents for each domain and favored the reduction of the following quantity.

$$S_{dom_i}(p) = \sqrt{\frac{1}{N}\sum_{h=1}^{D_{dom_i}} N_h \,(\overline{p}_h - \overline{p})^2} \qquad (1)$$

that is, the variability of the average of mean response propensity within the groups of a given type of domain.

A set of respondents is "representative" with respect to a domain if the previous one is null (or very low). An alternative form of (1) is the following:

$$R_{dom_i}(p) = 1 - 2S_{dom_i}(p) \qquad (2)$$

The latter indicator is in the range [0,1]:

- if R_{dom_i} is close to 0 then the sample is NOT representative for the dom_i and action is required;
- if R_{dom_i} is close to 1 then the sample is representative for the dom_i domain and no corrective action is required.

To raise the value of R_{dom_i} the following procedure was adopted:

1. initially, attention was focused on dom_1;
2. the response rate and the size in terms of enterprises have been calculated for each of the D_{dom_1}=2,626 sets that make it up;
3. these sets were then sorted:
    a. by ascending order of response rate (i.e. sorting begins with the domains with the lowest response rate);
    b. groups with the same response rate, by descending order of the their size (i.e. for groups with the same response rate the ordered list begins with the groups that contain more enterprises);

4. on the basis of this order, a reminder priority has therefore been assigned to each group (from 1 to the number of sets D_{dom_i} that constitute the partition of dom_i, therefore in the case of dom_1 from 1 to 2,626). Let $prior_{di}$ the priority assigned to the dom_i domain. In this way, 1 indicates the set of units on which the highest priority of recalling must be given;
5. steps 1-4 were then replicated also for the dom_2 and dom_3, giving $prior_{d2}$ and $prior_{d3}$;
6. each unit was then associated with the three priorities ($prior_{d1}$, $prior_{d2}$ and $prior_{d3}$) on the basis of the sets to which they belong with respect to the three types of domain;
7. each non-responsive unit was then associated with its highest priority $prior_{max}$=min($prior_{d1}$, $prior_{d2}$, $prior_{d3}$);
8. the selection of the units to be solicited has obviously taken place among the non-responding units on the basis of the highest priority;
9. since the same level of priority can derive from several domain types and the total number of reminders had to be less than 50,000, it was decided that to solicit up to 50% of non-respondent units. Such units were divided proportionally to the number of non-respondents in each domain for each priority level. Thus, for example, if the non-responding units with priority 3 were 150 of which 30 from the first domain 70 from the second and 50 from the third, 75 were included in the list to be solicited, of which 15 selected from the first domain, 35 from the second and 25 from the third.

The described procedure was applied twice during data collection period so that two different lists of units to be solicited were identified and used in separated waves.

3.3 Response rate analysis

The strategy for targeted solicitation involved in total approximately 60,000[6] units, counting the units included in the two waves, out of the total 278,402 units involved in the multi-purpose qualitative survey. Approximately two third of those extra solicited units were not registered into the Business statistical portal and more than half (around 32,000 units) were not registered and have less than 10 employees.

The solicitation activity started in February the 5th and the response rate, after 10 weeks of data collection, was at that time equal to 21.6 percent. Thus, there were still 8 weeks of data collection to go, as the end of data collection was fixed for the end of March, the 31st.

As already stated in § 3.1, targeted group were reached by extra reminders in two different waves, nevertheless the results - in terms of response rate obtained by the units involved in the two waves - are analyzed jointly in the following.

Looking at relationship between the response rate (*rr*) and the registration into the Portal (Table 3), it is possible to assess that *rr* it is always higher for registered units than for non-registered ones.

Moreover, considering the increment of response rate (percentage points – *pp*) registered between the day in which the targeted solicitation started (February the 5th) and the end of data collection, it is possible to see that it was always higher in the targeted group than in the non-targeted one (Table 3). Such difference was around 2.8 *pp* on average. Targeted solicitation showed to be more effective with units having higher propensity to participate to surveys, and thus the gap between the two groups increased with the increasing

[6] Total number of units included in the solicitation activity turned out to be different from theoretical numbers indicated in paragraph 3.1, as there were units involved in both waves, units not reachable for lack of contacting information, questionnaire sent before the first solicitation reminder reached them.

dimension of units involved, being around 20 and more *pp* per registered units with at least 250 employees and reaching the minimum for registered units with 10 to 20 employees (13.6 *pp*).

Looking at the effect on units not registered, it is possible to observe that there is an inversion of this particular trend as for medium to large non-registered units (with at least 10 employees) the gap between target and non-targeted group was lower (1.9 *pp*) than the one realized in smaller units (with less than 10 employees - 4.8 *pp*).

Table 3 – Response rate of units included - or not - in the targeted reminder group, per date, class of employees and registration state – Year 2022

Type of unit	Sample units involved		Response rate before solicitation (February 5th) (%)		Final response rate (%)		Increment recorded from February 5th till end of data collection (%)	
	Targeted reminder		Targeted reminder		Targeted reminder		Targeted reminder	
	without	with	without	with	without	with	without	with
500+ registered	1,311	301	23.5	2.7	94.4	93.0	70.9	90.3
250-500 registered	1,886	396	27.8	3.8	92.1	90.7	64.3	86.9
20-250 registered	55,028	7,493	29.2	3.4	77.1	68.0	47.9	64.6
10- 20 registered	25,672	4,081	33.5	3.6	76.8	60.5	43.3	56.9
10+ NOT registered	25,373	7,021	12.7	1.6	30.2	21.0	17.5	19.4
less than 10 registered	35,882	8,171	40.2	4.1	78.4	57.9	38.2	53.8
less than 10 NOT registered	73,573	32,214	20.9	2.4	40.2	26.5	19.3	24.1
Total	218,725	59,677	26.8	2.8	59.7	38.5	32.9	35.7

Source: Business statistical portal for state of registration and Asia 2021 for number of employees

In particular, analyzing better the characteristics of non-registered units, it is possible to verify that the ones newly introduced in the Statistical portal have a higher response to survey, whereas the ones already involved in previous surveys - but not participating to them - show to be the most reluctant ones. In fact, this group registered the lower response rate (8.3 *rr* for non-targeted units and 10.7 for targeted units) among the considered groups and the lower increment (+2.4 *pp*) due to the targeted reminder (Table 4).

Table 4 – Response rate of units included - or not - in the targeted reminder group, per date and registration state – Year 2022

Type of unit	Sample units involved		Response rate before solititation (February 5th) (%)		Final response rate (%)		Increment recorded from February 5th till end of data collection (%)	
	Targeted reminder		Targeted reminder		Targeted reminder		Targeted reminder	
	without	with	without	with	without	with	without	with
Old in the Portal - *NOT Registered*	25,979	12,088	7.4	1.2	15.7	11.9	8.3	10.7
Old in the Portal - *Registered*	119,779	20,442	33.3	3.7	77.9	63.3	44.6	59.6
New in the Portal	72,967	27,147	22.9	2.8	45.5	31.6	22.6	28.8
Total	218,725	59,677	26.8	2.8	59.7	38.5	32.9	35.7

Source: Business statistical portal

At territorial level, the northern regions are the ones that reached the highest response rate (Table 5). Among those, Lombardia is the one where the targeted units scored the highest increment (+19.9 *pp*) compared to non-targeted units. Among the southern regions, Molise showed the best reaction to targeted reminder (+11 *pp*) whereas Puglia and Sicilia scored the lowest values (+4.9 and +5.2 *pp* respectively).

Table 5 – Response rate of units included - or not - in the targeted reminder group, per date and region – Year 2022

Regions	Sample units involved		Final response rate (%)		Increment recorded from February 5th till end of data collection (%)	
	Targeted reminder		Targeted reminder		Targeted reminder	
	without	with	without	with	without	with
Piemonte	17,366	1,769	63.2	48.5	36.1	43.9
Valle d'Aosta/Vallée d'Aoste	993	271	61.3	44.6	34.0	42.0
Lombardia	40,651	3,927	66.9	62.9	39.0	58.9
Trentino-Alto Adige/Südtirol	5,653	580	61.4	51.9	29.2	44.7
Veneto	22,029	1,681	64.7	54.2	37.3	50.0
Friuli-Venezia Giulia	6,491	605	63.7	55.7	35.5	52.1
Liguria	5,849	1,641	59.4	43.4	32.6	40.2
Emilia-Romagna	23,365	1,445	64.8	52.9	36.8	48.1
Toscana	19,143	3,772	59.4	43.7	32.2	40.1
Umbria	3,694	1,269	58.6	43.2	32.6	40.5
Marche	9,283	2,023	59.7	44.9	33.4	41.7
Lazio	11,698	5,188	57.6	42.0	32.7	39.5
Abruzzo	5,321	3,248	53.2	36.1	27.7	33.4
Molise	1,328	1,627	45.6	32.4	18.9	29.9
Campania	10,941	7,140	48.3	32.5	24.4	30.0
Puglia	11,354	4,569	52.4	34.6	27.2	32.1
Basilicata	1,981	1,348	50.9	33.7	25.6	31.8
Calabria	5,063	4,607	39.4	27.6	17.4	25.8
Sicilia	10,767	9,232	45.5	28.2	21.1	26.3
Sardegna	5,755	3,735	50.2	33.8	24.5	31.6
Total	218,725	59,677	59.7	38.5	32.9	35.7

Source: Business statistical portal for state of registration and Asia 2021 for territorial information

3.4 Data quality analysis

The effects of the reminders carried out in the three groups of investigation (*dom₁, dom₂, dom₃*) were estimated using the q_2 index given below at three different time periods: before the first reminder (Reminder 0), after the first and before the second reminder (Reminder 1) and at the end of the survey (Reminder 2).

This indicator is an alternative to R-indicators and is used to estimate parameters such as population means and totals (Bianchi et al., 2016).

After calculating the response rate in the three groups for the three time intervals, the mean and standard deviation were calculated.

If we denote the standard deviation as *S* the mean as μ, and the domain *i* at time t as dom_{it}, we can define q_2 as:

$$q_{2dom_{it}} = S_{dom_{it}}/\mu_{dom_{it}} \qquad (1)$$

$for\ i\ \ 1{:}\,3; for\ t\ 1{:}\,3.$

Equation 1 represents the coefficient of variation of the response rate. The objective is its progressive reduction following the reminders, indicating a reduction in estimation bias.

Table 6 shows the values of q_2 for the domains before and after the two reminders and their respective percentage differences.

Table 6 - q_2 values per time interval, percentage differences and group

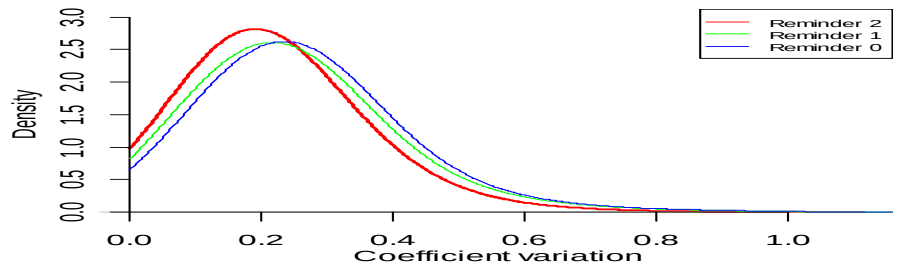
Groups	Reminder 0	Reminder 1	Reminder 2	Reminder 1/ Reminder 0	Reminder 2/ Reminder 1
	q2	q2	q2	Var. %	Var. %
Dom 1	0,63	0,56	0,33	-11,33%	-40,48%
Dom 2	0,83	0,71	0,40	-15,00%	-44,06%
Dom 3	0,53	0,44	0,27	-16,40%	-39,12%

Table 6 clearly shows that the reduction in q_2 was significant following the reminders in all three groups. In particular, dom_2 showed the greatest reduction (-44.06%), while dom_3 had the smallest decrease (-39.12%).

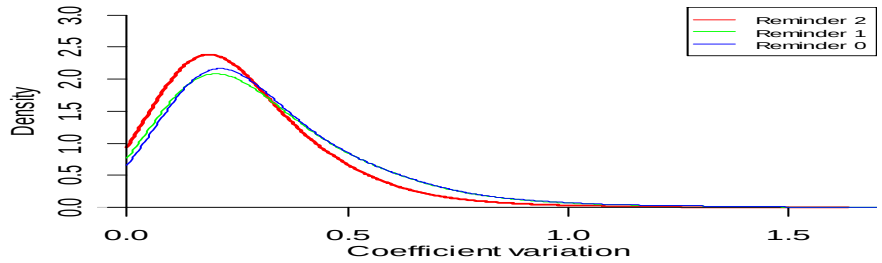
The following figures show the density estimation of coefficient variation for the two reminders, for the three groups (dom1, dom2, dom3).

The shift of the curves to the left side, shows the decrease in the coefficients of variation of the response propensities in the three groups.

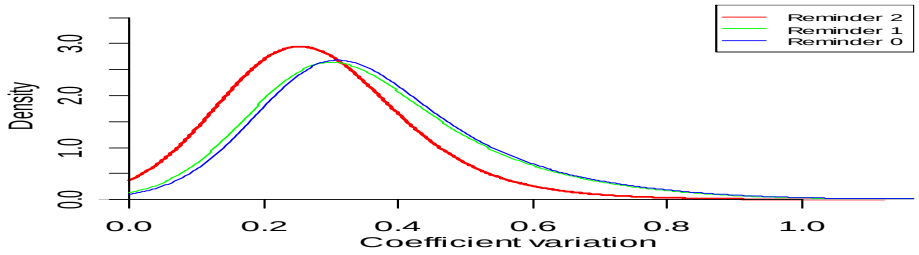
Graph1 - Density estimation of coefficient variation by dom1



Graph 2 - Density estimation of coefficient variation by dom2



Graph 3 - Density estimation of coefficient variation by dom3



4. Conclusions and future developments

The multi-purpose qualitative survey represents a unique tool for collecting relevant information on companies, otherwise not available in such an organic and timely manner. However, in general and in perspective, it is also necessary to reduce the overall statistical burden for companies, which are annually subjected to numerous (mainly quantitative) surveys. In this context, planning an efficient strategy of targeted reminders aimed at collecting relevant information while minimizing returns to non-responsive units appears to be the most effective way to obtain quality results.

The conclusions reached in the context of this work show the need to set up, in the data collection phase, a constant monitoring of the response rate trend with reference to specific domains. In fact, this activity makes it possible to obtain appreciable results precisely with the aim of correcting any distorting phenomena in real time and therefore guaranteeing a high level of quality in the final results.

The analysis carried out shows that an adaptive approach to the data collection strategy, and notably to the reminder strategy, represents an important tool for increasing the efficiency of the data collection process, reducing the estimate bias and, consequently, optimizing the available resources. In fact, the results obtained show a clear reduction in the variability of the propensity to respond in the three identified domains, based on the application of the R-indicators methodology, which is reflected in a reduction of the bias of the estimates.

Therefore, in the light of the results obtained for the multi-purpose qualitative survey of the Permanent Business Census, within the framework of the centralized data collection model adopted by Istat, it can be stated that the methodology can be extended to all other structural surveys on businesses. That will make it possible to respond to the increasingly stringent needs to optimize utilized resources by guaranteeing adequate quality standards and, at the same time, reducing the statistical burden on respondents.

Finally, the need to plan the cooperation and interaction between methodology and data collection in different phases of the statistical activity in direct surveys is evident so that the involvement of the units contacted in the context of the surveys can be reduced as much as possible.

References

Bellini G., M.C. Casciano, S. Filiberti, M. Piaggesi, M. Rinaldi. *Towards the adoption of adaptive contact strategies of units involved in business surveys*. UNECE Expert Meeting on Statistical Data Collection – Towards a

New Normal?. 26 to 28 October 2022, Rome, Italy. https://unece.org/sites/default/files/2022-10/DC2022_S3_Italy_Bellini%20et%20al_AD_0.pdf

Bellini G., Monetti F., Papa P. *The impact of a centralized data collection approach on response rates of economic surveys and data quality: the Istat experience*. Statistics and Economy Journal Vol. 100 (1) 2020 ISSN 1804-8765 (Online) *ISSN 0322-788X (Print).*

Bianchi, N. Sholomo, B. Schouten, D. da Silva, C. Skinner (2016), *Estimation of response propensities and R-indicators using population-level information*, 21, CBS, Discussion paper.

Eurostat. *Handbook for Monitoring and Evaluating Business Survey Response Burdens*. Luxembourg, 2003. https://ec.europa.eu/eurostat/documents/64157/4374310/12-handbook-for-monitoring-and-evaluating-business-survey-resonse-burden.pdf/600e3c6d-8e8d-44f7-a8f5-0931c71d9920

Istat. *Censimento permanente delle imprese 2019: i primi risultati*. Roma, 2020. https://www.istat.it/it/files//2020/02/Report-primi-risultati-censimento-imprese.pdf
https://www.istat.it/it/archivio/238337

Istat. *Situazione e prospettive delle imprese dopo l'emergenza sanitaria Covid-19 Statistica report*. Roma, 4 febbraio 2022. https://www.istat.it/it/files/2022/02/REPORT-COVID-IMPRESE_2022.pdf

Schouten B., M. Calinescu and A. Luiten. What are adaptive survey designs? In *Survey Methodology*. Volume 39, Number 1. Statistics Canada. Code 12-001-X. June 2013.