Received: 14 November 2022 | Revised: 10 February 2023 | Accepted: 28 March 2023

DOI: 10.1002/asmb.2761

RESEARCH ARTICLE

WILEY

The Italian quick survey on the effects of the COVID-19 health emergency on businesses: Sampling strategy and data editing

M. C. Casciano[1] | **R. Varriale**[1,2]

[1]Istat - Italian National Institute of Statistics, Rome, Italy

[2]Istat - Italian National Institute of Statistics, Sapienza University of Rome, Rome, Italy

Correspondence
R. Varriale, Istat - Italian National Institute of Statistics, Sapienza Università di Roma, Roma, Italy.
Email: varriale@istat.it

Abstract

During the first phase of the COVID-19 pandemic, Istat performed the quick survey "Situation and perspectives of Italian enterprises during the COVID-19 health emergency," with the aim of assessing the economic situation and the specific actions adopted by businesses to reduce the economic impacts of the emergency. To ensure the continuity in the information flow and to analyze the temporal evolution of the observed phenomena, the survey has been repeated in three different waves. The outcomes of each wave was released just after 2 months from the launch of the survey. The present work analyses the characteristics of the sampling strategy and describes the complexity of the data editing process, in the case of a survey planned to produce estimates able to ensure an acceptable level of accuracy in the maximum timeliness.

KEYWORDS

COVID-19, data editing, sampling strategy, statistical survey

1 | THE CONTEXT: THE ITALIAN QUICK SURVEY ON THE EFFECTS OF THE COVID-19 HEALTH EMERGENCY ON BUSINESSES

The health emergency that occurred around the world in 2020 had dramatic effects on businesses. In Italy, many organizations tried to analyze and evaluate these effects. Just to give some examples, the Bank of Italy has dedicated an entire part of its website to this topic.[1] Confindustria, that is the main association representing manufacturing and service companies in Italy, with a voluntary membership of more than 150,000 companies of all sizes, employing a total of 5,382,382 people, has launched a survey through an online questionnaire to Italian associated and nonassociated companies.[2] Participation in the survey was very high: 6,000 companies filled out the questionnaire and, of these, 4,000 were used for analysis purposes. Confcommercio-Imprese per l'Italia, the Italian General Confederation of Enterprises, Professional Activities and Self-Employment, that is the largest company representation in Italy, associating over 700,000 companies, also studied the topic.[3]

The Italian National Institute of Statistics, Istat, conducted the survey "Situation and perspectives of Italian enterprises during the COVID-19 health emergency" between 8 and 29 May, 2020, with the aim of collecting assessments directly from companies on the effects of the emergency health and economic crisis on their business.[4] The final aim was

M. C. Casciano and R. Varriale contributed equally to this work.

The content of the article is due solely to the authors and does not represent in any way the view of ISTAT on the subject.

to provide citizens, economic operators and public decision-makers with high quality and timely statistical evidence on how Italian companies were experiencing this difficult period, with reference to the economic and financial impact, and employment. The number of companies in the sample was 90,461. After this experience, the survey has been repeated in other two waves to ensure the continuity in the information flow and to analyze the temporal evolution of the observed phenomena. The survey performed between 23 October and 16 November, 2020 aimed at updating the information collected in the first edition and allowing for new assessments on the effects of the pandemic on business activities and their perspectives.[5] The survey performed between 16 November and 17 December, 2021 updated the information collected in previous editions by measuring the behavior and strategies of enterprises almost two years after the beginning of the pandemic.[6]

All the waves of the survey "Situation and perspectives of Italian enterprises during (and after) the COVID-19 health emergency" were characterized by the need to produce estimates able to ensure an acceptable level of accuracy in the maximum timeliness: the outcomes of each wave were released after two months from the launch of the survey. This work aims at describing the methodological characteristics of the sampling strategy and the complexity of the data editing process for this survey.

The article is organized as follows. Section 2 summarizes the methodological innovations introduced in the survey; Section 3 describes the sampling strategy, and Section 4 the questionnaires and data collection. The description of the editing and imputation phase and of the estimation procedure are in Sections 5 and 6. Section 7 concludes the work.

2 | INNOVATIVE ASPECTS OF COVID SURVEY

The survey "Situation and perspectives of Italian enterprises during (and after) the COVID-19 health emergency" is a sample survey on enterprises. The survey is composed by three waves, that have been run in May 2020, October and November 2020, and November and December 2021. The adopted sampling design is a two-phase stratified random sampling. Due to the special nature of the survey, a panel of economic units to be surveyed at subsequent times has been selected, aiming to submit them partly different questions, to analyze the evolution of the behavior and reactions of enterprises in the short—medium term with respect to the health emergency.

Covid survey presented important innovations both in the sampling strategy and in the contact plan of the (potential) respondents.

As for the sampling strategy, a two-phase sampling design has been implemented (see Sections 3.1 and 3.2). This is a complex sampling design, in which there are two (linked) surveys: the first phase survey has, among other things, the objective of collecting auxiliary information useful for defining the second phase sample. This auxiliary information is not available in the population frame and is correlated to the target variables of the second phase sample. In the survey "Situation and perspectives of Italian enterprises during (and after) the COVID-19 health emergency," instead of using the Istat business register as the list of units from which to select the sample, it was used the set of enterprises that had responded to the multipurpose survey related to the permanent census on businesses (PCE), whose results was just released on 7 February, 2020.[14] Therefore, PCE represented the first phase survey, and the Covid survey constituted the second one. The adoption of this strategy was possible because the unit of analysis for both surveys is the enterprise (legal unit). Furthermore, the consistency between PCE and survey's target populations offered the advantage for the survey sampling strategy that PCE provided updated information both with respect to structural information on enterprises and to qualitative information, otherwise not available, especially for small and medium-sized economic units: PCE provided a statistical information framework on enteprises' behavior, organization, management, human resources, relations between companies, internationalization, technologies and investments in innovation, environmental sustainability, and corporate social responsibility. Another innovation introduced in the estimation phase of the sampling strategy is the massive use of paradata via the application of random forest algorithms, with the purpose of reducing bias in final estimates due to the nonresponse (see Section 6.1).

As for the innovations related to the contact of the sampled units for Covid survey, for the sake of timeliness, it was decided to contact those enterprises owing a Certified Electronic Mail (PEC) and already registered on Istat Enterprise Portal (https://imprese.istat.it/), instead of sending the questionnaire to the whole set of the eligible units (i.e., the whole set of PCE respondents). As a further innovation, the information letter presenting the survey was sent via e-mail to the contacts of the enterprises thus identified for their direct involvement in filling in the PCE questionnaire even in the event of temporary closure of the enterprise. Infact, the information letter for Istat surveys is usually sent by email to the enterprise administrative department. Finally, the reminder plan aimed at contacting nonrespondent enterprises was also innovative: instead of a massive reminder towards all nonresponding units, a weekly monitoring plan was implemented,

TABLE 1 Planned estimation domains of permanent census on enterprises (first-phase sample)

	Structural variables		
Domain type	**Economic activity sectors Nace Rev.2**	**Size[a]**	**Geographical breakdown**
1	5 macro-sectors[b]	3 classes of employees	107 provinces
2	2 digits – divisions	3 classes of employees	21 regions
3	4 digits – classes	3 classes of employees	-
4	2 digits – divisions	4 classes of employees	-

Note: For some Divisions, the employees size class 3–9 was further splitted into 3–4, 5–9.

[a]Classes of employees (3–9; 10–19; 20 and more). Enterprises with at least 20 employees were considered as a take all stratum, the ones in the strata 3–9 or 10–19 have been sampled.

[b]The macro-sectors of economic activity considered are the following: industry; energy and water; construction; trade; other services.

aimed at identifying targeted nonrespondent units to be re-contacted. This reminder plan aimed at guaranteeing a prefixed level of accuracy for the final estimates, therefore groups of units belonging to strata with a low response rate and/or with a high sampling error of the provisional estimates have been detected and solicited. The final objective was to achieve an adequate number of respondents per stratum with respect to the established sampling error thresholds, both at national and estimation domains level.

3 | FRAME OF INTEREST AND SAMPLING DESIGN

3.1 | First phase sampling design

The first-phase sample, represented by the sample selected for the PCE survey, is a stratified random sample, with the strata obtained by cross-classifying the modalities of the structural variables defining the four types of domains of interest (see Table 1).

More precisely, the stratification underlying the four types of domain is the concatenation of NACE (economic activity code, 4 digits), classes of employees (four size classes: 3–4; 5–9; 10–19; 20 and more), and territorial breakdowns (107 Italian provinces). This stratification corresponds to the finest partition of the population that allows each domain to be obtained as a union of complete elementary strata. This way of stratifying implies that the domains of interest constitute "planned domains." The fact that all domains of interest are of the planned type has some considerable advantages from the point of view of survey design. In particular, this allows the sample to be allocated in the strata by predefining the expected precision levels of the estimates across all domains of interest.

In order to determine the optimal sample size according to some prefixed expected sampling errors of the target variables estimates on the planned domains, PCE final allocation has been determined via a multivariate multidomain approach.[7] In particular, the minimum sample size has been determined in order to ensure that the variance of sampling estimates of the variable of interest in each domain does not exceed a given threshold, in terms of coefficient of variation. The number of units to be selected in each stratum is defined as a solution of a nonlinear integer problem.

From a target population of 1,037,950 enterprises splitted into 87,596 elementary strata, the PCE survey's optimal sample size consisted of 285,414 units to be randomly selected without replacement and with equal probabilities. The overall number of respondents to the PCE survey was 182,666 enterprises.

3.2 | Second phase sampling design

From the 182,666 respondents to the PCE, the subset of 161,619 enterprises owing a Certified Electronic Mail and already registered on Istat Enterprise Portal has been identified; this subset represented the sampling list on which the second-phase sample has been allocated.

The allocation of units by strata has been determined with the aim of ensuring an adequate coverage of sample enterprises in each of the domains of interest for the Covid survey (see Table 2).

TABLE 2 Planned estimation domains on Covid survey (second-phase sample)

	Structural variables		
Domain type	**Economic activity sectors Nace Rev.2**	**Size[a]**	**Geographical breakdown**
1	2 digits – divisions	3 classes of employees	-
2	Sections	3 classes of employees	5 NUTS1

[a]3 classes of employees (3–9; 10–19; 20 and more).

TABLE 3 Sections of the survey questionnaire, waves I, II, and III

Section	Wave 1	Wave 2	Wave 3
1	Impact of COVID-19 until May 4th, 2020	Impact of COVID-19	The current situation
2	Health control measures	Health control measures	
3	Management and personnel policies	Management and personnel policies	Management and personnel policies
4	Impact of COVID-19 in the medium term		
5		Finance	Finance
6		Digitalization and technologies	Digitalization and technologies
7		Criticalities and strategic guidelines	Criticalities and strategic guidelines

The stratification adopted here with respect to the two domain types was defined by the concatenation of NACE (2 digits), classes of employees (three size classes: 3–9; 10–19; 20 and more), and territorial breakdowns (5 classes: North-East, North-West, Centre, South, and Islands). At the second phase level, the overall number of elementary strata is 6,872, while the number of distinct domains turns out to be 234 for the first domain type and 256 for the second type, respectively.

Given this stratification, three allocations have been performed; for each of them, overall and strata sample sizes have been determined under a specific hypothesis for the true value of the unknown population parameter p, that is, the frequency of the phenomenon of interest in the population. Given each assumption on p, the distribution of the expected coefficients of variation of the sampling estimate of p has been calculated for the two chosen domain types. In a first step, these coefficients of variation have been estimated under the hypothesis that all units of the second phase sample are respondents.

In the case, for example, of a hypothetical frequency p of the phenomenon of interest equal to 30%:

1. The first allocation (about 60,000 sample units) would allow obtaining, for 75% of the study domains, in both types of domains considered, an estimate of p with a sampling error of less than 23%;
2. The second allocation (around 80,000 sample units) would allow obtaining, for 75% of the study domains, in both types of domain considered, sample errors of less than 19%;
3. The third allocation (approximately 100,000 sample units), would make it possible to obtain, for 75% of the study domains, in both types of domain considered, an estimate of p with a sampling error of less than 16%.

The allocation of the sample units in the strata, which are multi-domain in nature, is made consistent with an approach that takes into account the phenomenon of nonresponse to the PCE survey (first-phase sample), in which a nonrandom MRT mechanism was observed, mostly determined by the size and geographical location of the enterprise. Therefore, in order to limit potentially distorsive effects due to the phenomenon of self-selection of respondents, a sampling with inversely proportional probability to the first-phase response rate was performed within the second-phase strata. The total number of enterprises in the sample was 90,461.

4 | QUESTIONNAIRES AND DATA COLLECTION

At each survey occasion, the survey questionnaire consists of different sections, as summarized in Table 3.

During the first wave, the sections of the questionnaire were four, with 20 questions, some of which with multiple choices. As this was the first edition of a new survey and, above all, given the uncertainty of the behavior of companies determined by the exceptional nature of the current situation, an open response method was used for some questions to identify any behavioral profiles not covered in other items. The survey questionnaire of the second wave consists of six sections. Compared to the first edition of the survey, the questionnaire has been designed to satisfy both the needs of tracing the evolution of enterprise behavior with respect to specific issues (Sections 1–3), and with the aim of intercepting the organizational solutions adopted by companies and planned for the future short term (Sections 4–6). The survey questionnaire of the third wave consists of five sections. Also in this occasion, the questionnaire has been designed to meet both the needs of tracking the evolution of business behavior with respect to certain issues (Sections 1 and 2), both with the aim of intercepting the organizational solutions adopted by companies and planned in the future short term (Sections 3 and 5).

The companies were invited to participate in the survey and to fill in the survey form via the Istat Enterprise Portal, through a computer assisted web interview technique. According to this approach, the questionnaire is filled in directly by the contact person of the company to whom any inconsistencies and/or incompleteness are reported immediately during the compilation of the form.

5 | THE NON-SAMPLING ERROR: THE EDITING AND IMPUTATION PHASE

Despite all possible precautions, data collected though a survey inevitably contain errors, determining a difference between the measurement of the phenomenon (estimate) and the real phenomenon. To produce statistical output of sufficient quality, statistical institutes carry out an extensive process of checking the data and performing amendments. This process of improving the data quality for statistical purposes, by detecting and treating errors, is referred to as statistical data editing.[8]

During the entire survey process, from the design to the data collection phase, various types of errors may occur that are not attributable to the sampling strategy adopted, but to the set of operations necessary to carry out the survey itself. These errors are commonly called non-sampling errors to distinguish them from sampling errors, due to the sampling design and estimator used. Different types of non-sampling errors can be identified: non-admissible values of the individual variables and logical inconsistencies between information in more variables.

Non-sampling errors may also appear as item (partial) or unit (total) missing values. The former means that some statistical units did not provide information for one or more items, while there is total nonresponse when there are statistical units for which it was not possible to obtain all the information for various reasons: errors of the list, refusal to collaborate in the survey or impossibility of finding the sampled unit. One of the main effects of total nonresponses is the noncoverage of the population under study, which may produce serious bias in the estimates when the nonrespondents systematically differ from the respondents. In sample surveys, the total lack of answers cause, in addition, a reduction in the sample size and, therefore, the increase of the relative sampling error of the estimates. Also for partial nonresponse, the reduction of the sample size makes estimates of the quantities of interest less efficient, and distortions in the parameters of interest are possible. However, the effects of partial nonresponse are not serious like those of the total nonresponse because in this case there is some information about the interviewee. Different treatments are usually used for partial and total nonresponse. In case of total nonresponse, information-based weighting methods obtained from the sample design can be performed adopting mainly two techniques based on a "direct approach" in the case of a correction for elementary units, or by strata, computing some "propensity stratification weighting" (see Section 6). For partial nonresponse, instead, methods of imputation are used, which consist in replacing the missing values with values appropriately chosen by using the available information on the statistical unit in question. The description of the methods of imputation used in the survey "Situation and perspectives of Italian enterprises during the COVID-19 health emergency" is in the following paragraph.

After the data collection phase, out of the total number of firms in the sample, the percentage of firms that completed the questionnaire in the three occasions was 46.9 (I wave), 44.6 (II wave), and 46.1 (III wave). The overall response rate of the survey is lower than that of PCE, equal to 75.4% among the units in the take-all strata (enterprises with 20 or more than 20 employees) and 59.9% among the randomly selected units (enterprises with less than 20 employees).[9] This phenomenon can be explained by the fact that the Covid survey, unlike the PCE, has no obligation to answer. Furthermore, the data collection period is very short and the current economic situation is peculiar. It is also important to notice that the overall response rate of the Covid survey is similar to that one of other Istat surveys on enterprises having similar

TABLE 4 Absolute and relative frequency distribution of the respondents, by type of questionnaires, wave III

Response type	Absolute frequency	%
Questionnaires sent	41,634	88.52
Questionnaires not sent, but complete	52	0.11
Questionnaires not sent, with a low number of missing values	274	0.58
Questionnaires not sent, «Important variables» missing	4763	10.13
Questionnaires sent, but «important variables» missing or Questionnaires duplicates	9	0.02
Questionnaires not sent, with a high number of partial missing values	300	0.64
TOTAL	47,032	100.00

size. Just to give an example, the Survey on the Enterprise Accounting System (EAS) has two modules: one for the Large enterprises (250 or more than 250 employees, EAS–LE), and one for the small and medium enterprises (less than 250 employees, EAS – SME). The EAS–SME includes a large amount of small to medium size businesses, having generally a lower propensity to survey participation. In particular, for the year 2020 the response rate of EAS–SME is equal to 43.6 and the response rate of EAS–LE is equal to 85.9.[10] The overall response rate is therefore 45.7%. Therefore, despite an information letter in which nonrespondents were encouraged to provide information useful for finalizing interventions for a more rapid exit from the crisis and the direct contact with the enterprise contact person, the response rate of the Covid survey is in line with that observed for other structural surveys on enterprises. This result is still evaluated as positive, considering the particular economical contingency and the fact that there is no obligation and no penalty for nonresponse.

The percentage of firms that provided a partial answer was 4.9 (I wave), 3.2 (II wave), and 1.8 (III wave). However, in all occasions, part of the partially completed questionnaires showed a "too high" number of partial nonresponses, so they were assimilated to unit nonresponse.

Table 4 shows the analysis of nonresponse for the third wave of the survey. To answer the questionnaire, enterprises had to register on the portal, access the questionnaire, answer the questionnaire and then submit it. Missing values were not allowed. The results in terms of the number of valid questionnaires were very satisfactory. Just to give an example, during the third wave, out of the 47,032 collected questionnaires, more than 88% have been sent by the firms and do not present any missing value. Furthermore, other 52 firms completed the questionnaire even if they did not send it. Among the other cases, only 274 questionnaires have been considered for imputation as they present item nonresponse for few items, while the others have been discarded and assimilated to unit nonresponse. For 9 units there were problems in the data collection phase: these are duplicates or there are missing values even if questionnaires have been sent by the firms.

5.1 | Treatment of non-sampling error

Although the electronic questionnaire usually guarantees a good quality of the collected information, the missing values to one or more questions, even not admitted, can sometimes arise due to problems incurred during the survey compilation, and in any case not predictable. Similarly, there can be some inconsistencies between information in more variables: errors in the questionnaire "paths" can be attributed to any malfunctions of the software or hardware components of the data collection instrument, or errors in the design phase of the questionnaire itself.

An editing and imputation plan with a probabilistic method acts at the level of a single unit (record) to identify and to correct random errors that occurred during the data collection process. These errors include, for example, the lack of answers to one or more items, the non-admissible values of each item and the logical inconsistencies between the information found that is not attributable to systematic causes.

For qualitative variables, errors are usually identified by means of a set of edit rules, also known as edits or checking rules, that indicate conditions that should be satisfied by the values of single variables or combinations of variables in a record.[8] The rules defined by the survey experts constitute the set of explicit rules. If a record does not satisfy the condition specified by an edit rule, the edit rule is activated, and the record needs to be treated. On the contrary, a record that does not activate any rule is exact with respect to the set of rules and therefore does not need to be corrected.

Once defined, the explicit edits are used to localize the errors in data. In automatic editing of business survey data, the error localization problem for random errors is usually solved by applying the Fellegi-Holt paradigm, which states that a record should be made consistent by changing the fewest possible items of data.[11] Once the erroneous values have been detected, they are replaced with other values by means of imputation.

In Istat, a common software used for editing and imputing data is CONCORDJava (Data Control and Correction with Java interface).[12] In particular, the module SCIA (System for Automatic Control and Imputation) is an automatic control and correction system for qualitative variables entirely developed in Istat according to the methodology and formalisms of Fellegi and Holt.[11]

In the localization phase, to identify for each wrong record which variables to modify to bring the record back to a situation of correctness with respect to the set of rules defined, the Fellegi and Holt method requires that also implicit edits are logically derived from explicit edits. Explicit and implicit edits constitute the complete set of edits, essential to ensure the final correctness of a record. As described, the system identifies the minimum number of variables to be modified. Once the erroneous values have been identified (localization phase), in the imputation phase these values are replaced with those of the corresponding variables of one or more records in the set of donors, by applying a joint imputation first and, if necessary, a sequential imputation after.[12] In joint imputation, the incorrect values are corrected simultaneously by attributing to them the values that the same variables assume in the donor record; in sequential imputation, the variables are corrected separately, identifying, if necessary, as many donors as the number of incorrect variables.

In the three waves of the survey "Situation and perspectives of Italian enterprises during (and after) the COVID-19 health emergency," the entire control and correction process has been divided into two independent subprocesses due to the high number of explicit edits: (i) enterprises that have closed and do not plan to reopen, (ii) enterprises that are partially or totally open, or even if they are closed, plan to reopen. Just to give an example, in wave III, companies that are partially or totally open, or even if they are closed, plan to reopen have 312 explicit edits and a complete set of 2,310 edits.

In the third wave of the survey, the editing and imputation process involved 41,960 statistical units. Out of these, 620 are closed companies that do not plan to reopen: these units do not have incorrect records with respect to the set of rules defined. The other 41,340 units are companies that: sent the questionnaire definitively (41,634 units), did not send the questionnaire definitively but did not show partial nonresponse (52 units), did not send the questionnaire definitively and have no partial response (274 units). Out of these 41,340 units, 279 have been corrected using SCIA: 266 by joint imputation and 13 by sequential imputation. This result shows a good quality of the observed information, even in the presence of missing or incompatible data. Table 5 reports the frequency distribution of number of variables that have been corrected. It is important to note that the questionnaire was composed by 26 questions, many of which with multiple choices representing the variables in the software. The peaks show common behaviors among the respondents: most of them, in fact, did not complete the last section of the questionnaire.

Summarizing, the overall percentage of records corrected by using SCIA is equal to 0.66 (279 out of 41,960), showing a good quality of the analyzed information. During the other two waves, the overall percentage of records corrected by using SCIA is equal to 2.50 (I wave) and 1.26 (II wave), which shows an improvement in the quality of the survey process over time.

6 | ESTIMATION PROCEDURE

Covid survey's final estimates could be strongly affected by bias, depending from the level of the nonresponse on both stages of sampling. A second source of bias could be the undercoverage of the population of interest, due to the necessity of contacting the enterprises owing a Certified Electronic Mail and registered on Istat Portal for reason of timeliness. Both these sources of bias on the final estimates have to be corrected by the weighting procedure.

The final weights assigned to the sample units can be obtained by means of a complex procedure that starts with the calculation of the direct weight as the reciprocal of the probability of inclusion of each sample unit.

Two sets of adjustment factors are applied to the direct weights: the first, called nonresponse correctors, are determined by more or less complex techniques and are intended to correct, at least partially, the bias resulting from total nonresponse. The second set of adjustment factors, known as post-stratification factors, are subsequently computed to correct for the undercoverage of the list. The two sets of adjustment factors, together with their application to the Covid survey, are described in the next subparagraphs.

TABLE 5 Absolute frequency distribution of the edited records, by number of edited variables, wave III

Number of edited variables	Absolute frequency
1	18
2	7
3	1
4	1
5	4
6	3
7	12
8	14
9	2
10	2
11	3
12	3
13	1
14	1
15	14
16	3
17	146
18	8
19	3
20	6
21	16
22	3
23	6
24	2
Total	279

6.1 | Bias of the estimate's correction

Total nonresponse and the potential bias of survey estimates are a major problem, especially regarding business surveys, where (in contrast to social surveys) the different economic importance means that not all units contribute equally to survey estimates. Reduction of the accuracy of statistical estimates and bias effects are possible where there are strong dissimilarities between units participating in the survey and those not participating.

Estimates' bias can therefore be seen as the result of a correlation between surveyed phenomena, enterprise characteristics and the propensity of units to provide the required information. The knowledge of the random mechanism underlying this phenomenon (i.e., which ones among the set of unit's characteristics lead to a differentiation in the propensity of response) would be necessary and sufficient to overcome this problem.

Although this mechanism is not known, the statistical literature has developed several techniques having the aim of assigning an estimate of response propensity to each unit, starting from information available for both the set of responding and nonresponding ones. Such an estimate makes it possible to mitigate, or eliminate, the negative effects of nonresponse if the assumptions for estimating the response propensity, correctly describe the actual situation.

The response propensity model is used to calculate "corrective" coefficients to be associated with responding units only: the corrective coefficient is used to expand the role of responding units to also represent nonresponding units.

Response propensities, applied as the basis for survey nonresponse adjustments, can be made using two techniques: direct adjustment ("direct approach") or by stratum ("propensity stratification weighting"). In the first case, the adjustment factor to respondents' direct weights is calculated as the inverse of the estimated propensity to respond, while the second one assumes that units belonging to homogeneous strata of respondents have the same predicted response propensity.

In the case of direct estimation, as in the present case, there is a variety of techniques for identifying the pattern of response propensity, which is always expressed as a function of quantitative and categorical covariates. While discrete variable models, logistic or probit, have typically been employed for this purpose, in the of Covid survey, and also in PCE, an innovative method based on random forest (RF) was used. In recent years, RF methods are getting prevalent and commonly performed in statistical practice because of the advantages that the use of nonparametric ensemble models provide over linear models: convergence of the algorithm and handling of a large number of covariates, ability to capture nonlinearities in the data, overall performance of the ensemble classifier over linear models. In a nutshell, RF models are recursive classification methods based on known auxiliary information for both respondents and nonrespondents.

The information used in the case of Covid survey can be distinguished into two types: administrative information and paradata. The former, which typically come from administrative sources or statistical registers, describes units through variables such as: location, size, sector of economic activity, income statement, and organizational form. The paradata, on the other hand, describe the process of observing the unit through variables such as: interview technique, mode of contact, type of questionnaire, number of surveys in which the unit was involved and so forth.

The explanatory variables identified, that is, those ones used to define the profile of enterprises with respect to their propensity to respond, were several dozen (about 80, some of them, however, were strongly correlated each other).

The response propensity of each unit is expressed as the fraction of trees that had a respondent profile as an outcome out of the total 200 models. The results were satisfactory for the intended corrective purposes.

6.2 | Calibration procedure and sampling errors estimation

The second set of adjustment factors, known as post-stratification factors, are applicable when there are known totals of auxiliary variables correlated with the variables being surveyed; these factors have the property of making the final estimates more efficient than those based on direct weights alone, the greater the efficiency, the greater the correlation between the auxiliary variables and the surveyed variables; they also make it possible to mitigate the distorsive effect due to undercoverage of the list from which the sample is selected.

Post-stratification factors are determined by solving a constrained minimum problem: the function to be minimized is a distance function (appropriately chosen) between the direct weights (corrected for MRT) and the final weights, while the constraints are defined by the condition of equality over certain partitions (named calibration domains) between sample estimates of the totals of the auxiliary variables considered and the values, known from the population register, of the same totals. The distance function normally adopted is the logarithmic function, which ensures that the final weights are positive.

The final weight is finally obtained as the product of the direct weight by the correction factors.

For Covid survey, the auxiliary variables on which the condition of equality between known totals (called benchmarks) and respective sample estimates was imposed, were the average number of employees and the overall number of enterprises, both of them available at the elementary data level on the Business Register 2018—year of reference. The availability of the most up-to-date version of the Business Register (May 2020) made it possible to correct for the bias due to the phenomenon of under-coverage or duplication of units in the selection archive (Business Register 2018) at the calibration stage.

The partitions with respect to which the system of constraints was imposed on the known totals of the auxiliary variables, that is, the calibration domains, coincided with the dissemination breakdown of the estimates, as it can be seen in Table 6.

Regarding these partitions, perfect convergence was achieved between known totals from Business Register and sample estimates, with minimal distance between direct and final weights.

The final weights determined at enterprise level were multiplied by the values of the surveyed variables and then summed to obtain the estimates for the domain of interest.

Once available, the estimates for the target variable (which are mainly qualitative ones, meaning presence/absence of a characteristic on the respondent), their relative sampling errors in terms of coefficient of variations (CVs) have been

TABLE 6 Covid survey's dissemination breakdown

	Structural variables		
Domain type	**Economic activity sectors Nace Rev.2**	**Size[a]**	**Geographical breakdown**
1	2 digits – divisions		1
2	macro-sectors[b]	4 classes of employees	
3	macro-sectors[b]		21 regions
4		4 classes of employees	4 NUTS1

[a]Classes of employees (3–9; 10–49; 50–249; 250 and more).
[b]The macro-sectors of economic activity considered are the following: industry; energy and water; construction; trade; other services.

TABLE 7 Quantiles of estimated coefficients of variation (CV) distribution, for observed p-frequency levels of a qualitative target variable, by domain type of dissemination

***p*-freq**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**	**50%**
CV	Divisions of economic activity						
min	6.61%	4.55%	3.61%	3.03%	2.62%	2.31%	1.52%
median	21.65%	14.90%	11.82%	9.93%	8.60%	7.59%	4.97%
q_75	32.47%	22.35%	17.73%	14.90%	12.90%	11.38%	7.45%
q_90	45.66%	31.43%	24.94%	20.95%	18.15%	16.00%	10.48%
CV	Regions						
min	5.02%	3.46%	2.74%	2.31%	2.00%	1.76%	1.15%
median	10.91%	7.51%	5.96%	5.01%	4.34%	3.82%	2.50%
q_75	16.54%	11.39%	9.03%	7.59%	6.57%	5.80%	3.80%
q_90	24.45%	16.82%	13.35%	11.22%	9.71%	8.57%	5.61%
CV	Sector of economic activity * Classes of employees						
min	4.80%	3.30%	2.62%	2.20%	1.91%	1.68%	1.10%
median	7.00%	4.82%	3.82%	3.21%	2.78%	2.45%	1.61%
q_75	10.47%	7.20%	5.72%	4.80%	4.16%	3.67%	2.40%
q_90	24.42%	16.80%	13.33%	11.20%	9.70%	8.56%	5.60%
	Nuts1* Classes of employees						
min	5.49%	3.78%	3.00%	2.52%	2.18%	1.92%	1.26%
median	7.42%	5.10%	4.05%	3.40%	2.95%	2.60%	1.70%
q_75	11.09%	7.63%	6.05%	5.09%	4.41%	3.88%	2.54%
q_90	21.69%	14.93%	11.85%	9.95%	8.62%	7.60%	4.98%

calculated, in correspondence of the domains against which the estimates have been disseminated (see Table 7). Estimates resulted to be of good accuracy and consistent with the available external information.

7 | DISCUSSION

The three waves of the statistical survey "Situation and perspectives of Italian enterprises during (and after) the COVID-19 health emergency" were conducted by the Italian National Institute of Statistics between May 2020 and December 2021. The surveys provided citizens, economic operators and public decision-makers information on the effects of the

emergency health and economic crisis on Italian companies. In particular, the information collected made it possible to identify some behavioral profiles of Italian companies.[4–6]

In order to design and implement the statistical survey, important methodological efforts were done. The main feature of the survey is the timeliness of all phases of the statistical production process, from the design to the dissemination of information:[13] the outcomes of each wave were released after two months from the launch of the survey. Furthermore, the survey began during the first period of the pandemic, at a time of great crisis in the business world. This required an important and delicate phase of designing the survey. In particular, the results of the multipurpose survey related to the permanent census on businesses, which had just been released on 7 February, 2020,[14] were used to build the list of units from which the sample units have been selected. The adopted sampling design was a two-phase stratified random sampling. Due to the special nature of the survey, a panel of economic units to be surveyed at subsequent times (waves) has been selected. Response rate was higher than 44% in all three waves and was in line with that observed for other Istat structural surveys on enterprises. As discussed, this result is positive, considering the particular economical contingency and the fact that there was no obligation and no penalty for nonresponse. The percentage of questionnaires that needed to be corrected decreased from 2.5% to 0.66% in the three occasions. Covid survey's final estimates could be strongly affected by bias, depending from the level of the nonresponse on both stages of sampling. A second source of bias could be the undercoverage of the population of interest, due to the necessity of contacting the subset of enterprises already respondents to permanent census on businesses for reason of timeliness. The sampling strategy proved to be robust with respect to both sources of bias, providing estimates of good accuracy and consistent with the available external information.

In summary, the outcomes of the survey from a methodological point of view were of high quality, showing an excellent ability to react by Istat to new and timely information needs.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ACKNOWLEDGMENT

Open Access Funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

ORCID

M. C. Casciano https://orcid.org/0009-0006-6787-5372
R. Varriale https://orcid.org/0000-0002-2097-3833

REFERENCES

1. Bank of Italy. Covid-19 notes – 2021. Retrieved March 11, 2022. Available from https://www.bancaditalia.it/pubblicazioni/note-covid-19/2021/index.html?com.dotmarketing.htmlpage.language=1
2. Confindustria. Risultati relative all'indagine sugli effetti del Covid-19 per le imprese italiane; 2020. Retrieved March 11, 2022. Available from https://www.confindustria.it/notizie/dettaglio-notizie/indagine-effetti-Covid-19-imprese-italiane
3. Confcommercio-Imprese. IMPRESE E PANDEMIA: SCENARIO E NUMERI DELLA CRISI; 2021. Retrieved March 11, 2022. Available from https://www.confcommercio.it/-/imprese-pandemia-scenario-della-crisi%23:~:text=riduzione%20del%20reddito%20disponibile%3B,alle%20restrizioni%20alle%20attivit%C3%A0%20economiche
4. Istat. SITUAZIONE E PROSPETTIVE DELLE IMPRESE NELL'EMERGENZA SANITARIA COVID-19, 15 GIUGNO; 2020. Retrieved March 11, 2022. Available from https://www.istat.it/it/files//2020/06/Imprese-durante-Covid-19.pdf
5. Istat. SITUAZIONE E PROSPETTIVE DELLE IMPRESE NELL'EMERGENZA SANITARIA COVID-19, 14 DICEMBRE; 2020. Retrieved March 11, 2022. Available from https://www.istat.it/it/files//2020/12/REPORT-COVID-IMPRESE-DICEMBRE.pdf
6. Istat. SITUAZIONE E PROSPETTIVE DELLE IMPRESE DOPO L'EMERGENZA SANITARIA COVID-19, 4 FEBBRAIO; 2022. Retrieved March 11, 2022. Available from https://www.istat.it/it/files//2022/02/REPORT-COVID-IMPRESE_2022.pdf
7. Bethel J. Sampling allocation in multivariate surveys. *Surv Methodol*. 1989;15(1):47-57.
8. Eurostat. Memobust handbook. Theme: statistical data editing; 2014. Retrieved March 11, 2022. Available from https://ec.europa.eu/eurostat/cros/system/files/Statistical%20Data%20Editing-01-T-Main%20Module%20v1.0_1.pdf
9. Istat. IL PRIMO CENSIMENTO PERMANENTE DELLE IMPRESE. Istat: Letture statistiche – Temi; 2022. ISBN: 978–88–458-2061-8. Retrieved January 31, 2022. Available from https://www.istat.it/it/files//2022/03/Il-primo-censimento-permanente-delle-imprese_Ebook.pdf. doi:10.1481/Istat.Rapportoimprese.2021

10. Bellini G, Casciano MC, Filiberti S, Piaggesi M, Rinaldi M. Towards the adoption of adaptive contact strategies of units involved in business surveys. UNECE Expert Meeting on Statistical Data Collection, 26 to 28 October 2022, Rome, Italy; 2022. Retrieved March 11, 2022. Available from https://unece.org/sites/default/files/2022-10/DC2022_S3_Italy_Bellini%20et%20al_AD_0.pdf
11. Fellegi IP, Holt D. A systematic approach to automatic edit and imputation. *J Am Stat Assoc*. 1976;71(353):17-35.
12. Margarucci. CONCORD V. 1.0. Controllo e correzione dei dati. Manuale utente e aspetti metodologici. Istat - Strumenti e tecniche: Produzione libraria e centro stampa; 2004. Retrieved March 11, 2022. Available from https://www.istat.it/it/files/2011/03/manualeconcord.pdf
13. Eurostat. Memobust handbook. Theme: GSBPM: generic statistical business process model; 2014. Retrieved March 11, 2022. Available from https://ec.europa.eu/eurostat/cros/system/files/General%20Observations-06-T-GSBPM%20v1.0_1.pdf
14. Istat. RAPPORTO SULLE IMPRESE 2021. STRUTTURA, COMPORTAMENTI E PERFORMANCE DAL CENSIMENTO PERMANENTE. Istat: Letture statistiche – Temi; 2021. ISBN: 978-88-458-2068-7. doi:10.1481/Istat.Rapportoimprese.2021

How to cite this article: Casciano MC, Varriale R. The Italian quick survey on the effects of the COVID-19 health emergency on businesses: Sampling strategy and data editing. *Appl Stochastic Models Bus Ind*. 2023;1-12. doi: 10.1002/asmb.2761