Orietta Luzi, Istat

Ensuring statistical production and quality in the COVID-19 pandemic: the Italian experience

Stefano Falorsi[1], Orietta Luzi[1], Monica Scannapieco[1], Mauro Scanu[1]

[1] Istat, via Cesare Balbo 16, Roma

Abstract:
The COVID-19 pandemic has led to a dramatic loss of human life worldwide and posed unprecedented challenges to public health, economy and the labour market.
In order to respond to the new challenges posed by the pandemic on Official Statistics production, Istat put in place research and innovation activities that allowed to ensure governments and policy makers with reliable and continuous information in the response phase of the crisis. The methodological architecture of key social and economic surveys was revised to face the increased non response rates mainly deriving from the need to opt out of face-to-face/CAPI interviews due to social distancing rules. New data sources were explored to support current Official Statistics, and new surveys were designed to measure the impact of the COVID-19 emergency on Italian households and businesses, and to estimate the siero-prevalence propagation rate in the Country. Many of such innovations are now in the course of becoming structurally embedded in current survey processes, with the need for Istat to plan adequate investments for this purpose.

Keywords:
COVID-19; Official Statistics; Non response; Data collection; Non traditional data sources

1. Introduction:

Official Statistics have played an important role in the management of the COVID-19 pandemic, by ensuring governments and policy makers with continuous and reliable statistical information on society, economy and environment during the whole pandemic period, characterized by social distancing and lock-down phases. As many other National Statistical Institutes (NSIs hereafter), Istat was not actually ready to face the new challenges posed by the pandemic both in terms of ensuring the continuity and the quality of the current statistical production, and of measuring the impact of the pandemic on households and businesses, as the currently available information was not intended to be used to respond to such type of emergency needs.
In the “response phase” of the pandemic, which is still ongoing, Istat efforts have been primarily addressed on how to revise methodological and data collection strategies of the current survey processes, and how to possibly use additional sources of information (including Big Data), in order to compensate for low response rates and reduced data quality and filling-in information gaps, thus ensuring the continuity and the quality of current statistical production. In addition, new surveys have been designed and carried-out to assess the actual impact of the emergency on households and on enterprises, and to measure the siero-prevalence phenomenon in the Italian territory.
In this paper the most important methodological innovations introduced in Istat in the pandemic period are illustrated. The paper is structured as follows. In Section 1, the key innovation aimed at ensuring the regular release of the most relevant statistics with high reliability and quality levels are illustrated. In Section 2, the main methodological elements of the new surveys designed to measure the impact of the pandemic on households and businesses are

described. Section 3 contains an overview of some relevant experiences carried out in the area of using non-traditional sources of data to support current survey processes. In Section 4, some general conclusions are reported.

2. Ensuring continuity and quality of current statistical production:

As mentioned, the COVID-19 pandemic caused a general increase in the total non-response rate, with two major drawbacks: the observed sample is smaller, causing a general increase in the estimators coefficient of variation, and the non-respondent group could be non-ignorable for the target parameter under estimation. A specific Task Force has been set-up at Istat to face these problems and find methodological and technical solutions to guarantee the continuity and the quality of statistical production during the emergency. The Task Force was in charge of analysing all the incoming production processes and statistical releases, selecting the most urgent ones and evaluating, case-by-case, possible different actions. There were many surveys which require CAPI technique, such as HBS (Household budget survey), or a mix of CAPI and CATI techniques, such as Permanent Population Census, Labour Force Survey, just to cite some relevant examples. In line with the Italian Government rules on social distancing, all face-to-face interviews were stopped to guarantee security. In the following paragraphs, we mention some of the main statistical approaches that have been introduced in the current production processes based on the analyses carried out by Task Force.

Change in the estimation strategy – A usual approach in order to tackle total non-response consists of appropriately revising the estimation strategy, taking into account the need to change the data collection strategies. This implies revising the sampling design and using additional variables in calibration models. The most relevant survey where the estimation strategy has have been revised during the COVID-19 pandemic is the Labour Force Survey (LFS). It is important to recall that the Italian LFS adopts a mixed CAPI-CATI mode of data collection. In particular, according to the sample rotation scheme (2-(2)-2), households participate four times to the survey over a 15-months period: the first interview is always conducted by CAPI; the following ones are mainly conducted by CATI. During the 2020 lock-down, CAPI interviews were not allowed at all, and interviewers were asked to collect data only by phone (whether the phone number was available). Moreover, the company in charge of CATI interviews interrupted working until teleworking was set up (more or less all the lock-down period), in the meantime the interviews expected to be done by CATI interviewers were delivered to CAPI interviewers to be conducted by phone as well. These difficulties on data collection had a strong impact in the end of the first quarter 2020 (month of March), causing lower response rates, both in CAPI and CATI modes, and higher substitution rate (in CAPI) of the households. In order to reduce the impact of COVID-19 pandemic on data collection, a new sample to be used for the first wave in 2020Q2 was selected, composed by only households for which phone number was available. In this way, a phone number was available for all the households to be interviewed for the first time (and most of the households at the following waves). However, in order to analyse the possible bias deriving from this approach in 2020Q2 collected data, the distribution of the respondent sample was compared with previous quarters and with administrative sources. The results show in 2020Q2 higher frequencies of elderly people, Italian citizens, people not working in 2019, people having higher education level. A change in the estimation process was then necessary: besides the already used age and citizenship, further auxiliary variables related to the distribution of the individuals by educational level were introduced in calibration. The total was derived by the LFS estimates of the population by educational level in 2019 by: Nuts-2 regions; 4 age groups: 15-29, 30-49, 50+; 4 educational levels. As for administrative employment, the weighted sample finally reduced also the bias compared to the unweighted one. Due to the reduced sample size and the higher variability of final weights, higher sampling errors for the final estimates were estimated.

Introduction of a response adaptive design – Responsive adaptive design (Schouten and Shlomo, 2017) is a method that adapts data collection to characteristics of the survey target

population. This idea proved to be extremely useful in the lockdown months where, as already said, a high total nonresponse rate characterized many surveys. Ballin and Guardabascio (2020) use the responsive adaptive design approach in order to use the characteristics of the propensity to answer and suggest which non-respondents should be analysed more in detail. The core idea, already present in Shouten et al. (2009) describes the bias of the Horvitz-Thompson estimator of the mean of a variable Y as a function of the covariance between Y and the propensity to respond of each unit. If there is not auto-selection of the missing items, the covariance is null and the estimator is not biased. If the set of respondents seems to under- or over-represent some groups in a correlated way with Y, the idea is to reshape the set of units to collect assigning priority to those units that help fixing the problem. i.e. tend to reproduce the absence of auto-selection. This correction is made checking the R indicator, described in terms of a between variance of the average propensity to respond in groups and the overall average propensity to respond. This approach has been applied during the lockdown months in the Quarterly survey on turnover in the services. Data on the first quarter in 2019 and 2020 were compared (in 2019 respondents were 82,7% of the sample, while in 2020 respondents reduced to 46% of the planned one). In the two occasions, the indicator R was similar (0,787 in 2019 and 0,727 in 2020). Anyway, it is not appropriate to conclude that the situation in the two years is also similar. In fact, 2019 had a much smaller non-respondent ratio. Hence, it can be expected to have a maximum of 10% bias with an R indicator greater than 0,668 (as a matter of fact, the actual value fulfils this inequality). On the contrary, the non-respondents rate is much greater in 2020, and the same tolerance with respect of bias could be obtained only if R is greater than 0,816. Hence, the observed sample in 2020 is at risk of being biased. The authors, with essentially the same indicator, identified what NACE codes and geographic regions should be primarily included in the sample in order to correct the estimator from the bias effect due to auto-selected non-respondents. As a matter of fact, this approach proved to be a valuable tool that can be used in general, also in non-COVID periods and for other surveys.

Change in the imputation strategy – There are few cases in which the completely missing items have been imputed. An example is given by a survey that is traditionally not affected by large non response rates: Monthly survey on employment, working time, earnings and labour cost in large enterprises. Another example is given by Prodcom: in this case, data observed in 2018 have been projected in 2019 by means of generalized linear models (accuracy measures are provided by boosted regression trees). With the use of the *projection estimator* for imputing missing values, it was possible to enhance a sample consisting of only 35% of respondents, covering missing information on 60% of the planned sample. Another approach exploits additional sources of information in terms of aggregate data: e.g., as suggested by Eurostat regulations, VIES (VAT Information Exchange System) data on aggregate values on the exchange trade between member states per NACE code have been used in the Foreign trade survey in order to guide the imputation mechanism and improve overall estimates.

Change in time series adjustment – A number of key surveys in both households and economic area needed methodological support for time series treatment. In many cases (e.g. Foreign Trade), additive outliers have been included in the model at the end of the series, as “irregular” components that have the objective to represent the shock in specific months. In general, Istat followed the methodological guidelines provided by Eurostat (Eurostat, 2020).

The lock down and social distancing rules during the pandemic had an important impact on censuses, too. In particular, the direct surveys (Master Sample) of the Italian permanent population census were not carried out for 2020 round, and only estimates mainly arising from administrative registers, as population distribution by sex, age and citizenships and by educational level, are going to be disseminated. This experience will have strong impact on the design of the next 2022-2025 round of the census, as many innovations in terms of data collection strategy, use of administrative and non traditional sources, sampling and estimation

strategy will be revised taking into account the research work made to face the methodological issues posed by the pandemic limitations.

3. Measuring the impact of the pandemic on households and businesses:

In Spring 2020, several NSIs carried out sierological surveys of SARS-CoV-2. However, many of them were small or based on non-random sampling of participants (e.g., focusing on health-care workers or blood donors) and thus could not provide precise estimates of siero-prevalence by age groups in the general population. Additionally, some of these studies have used antibody tests with low sensitivity or specificity or have not reported the characteristics of the test siero-prevalence by age group in the general population. In April 2020, the Italian Ministry of Health and Istat, in collaboration with the Italian Red Cross the Regions that carried out the field operations, launched a nationwide, population-based, siero-epidemiological survey, aimed to estimate the extent of SARS-CoV-2 diffusion in the country (https://www.istat.it/it/archivio/246156). In particular, the survey aimed to evaluate together with the serum prevalence rate for SARS-CoV-2 in the population, the fraction of asymptomatic and subclinical infections. It was planned a nationwide sample of 150.000 individuals randomly selected from Istat's Population Register. In order to deal with expected high non-response rates - in a context in which sample substitution mechanisms were not recommendable - an oversampling rate of 25% was applied leading the final sample size to 195.000 individuals. The survey was aimed to produce a detailed snapshot of the phenomenon of interest in spring 2020 being representative of Italian population by pre-defined domains of interest. Indeed, technical literature on the epidemiological studies on SARS-Cov-2 epidemics shows how sampling selection procedures based on longitudinal sample and tracing rules may result effective in improving the efficiency of the final estimates (see, for example, the recent works by Alleva et al., 2020). The survey was conducted between May 25 and July 15, 2020. To the sample individuals, in addition to being subjected to a blood sample to carry out the serological test, were administered a short questionnaire aimed at detecting the presence of symptoms and risks factors. The main parameters of interest of the survey concern the rate of individuals according to their epidemiological status with reference to different sub-populations related to territorial and/or structural characteristics of the investigated population referred to as domains of interest. In particular, the primary territorial domains of interest are the Italian Geographical Regions and Autonomous Provinces of Bolzano and Trento, while the structural primary domains, within each geographical region, for the general population consist of age groups, and sex by large age groups. Furthermore, for working people sub-populations within each region, four economic activity macro-classes are considered. The percentage of respondents at National level was approximately 38% of the initial sample. Furthermore, only about 34% of the entire sample underwent the serological test, the target variable of the epidemiological study. For this reason, the construction of sampling weights was particularly accurate in order to try to mitigate the potential bias in the final estimates. A first weighting step was carried out to correct final estimates for non-response, by means of competing models for total non-response. A final calibration step was applied to non-response weights in order to take into account of known population totals by demographic and educational level. A validation phase on final estimates, which was carried out by using external data available from different geographical regions and for specific municipalities, has not evidenced particular bias.

In 2020, Istat also carried out two new surveys in order to measure the impact of the pandemic on households and enterprises. As far as households, a new survey was carried out two times (in spring and in winter 2020), in order to assess the social impact of COVID-19 emergency on households in terms of different topics, such as a day's diary under the coronavirus, respect for distancing rules, trust in health personnel, renunciation of medical treatment for fear of contagion; issues about depression and mental health; use of informal help networks; family workload on women (https://www.istat.it/it/files//2020/05/Reazione_cittadini_lockdown.pdf). Samples of 3,000 households were drawn from 2019 Master Sample of the Population Census. A new web survey was also carried out to measure the effects of the health

emergency on Italian enterprises, especially in terms of economic, employment and financial impact (https://www.istat.it/it/archivio/249361). The survey was carried out two times (in spring and autumn 2020) on a sample of about 90,000 enterprises drawn from the ones that responded to 2018 permanent census on enterprises. The imputation strategy, consistently with the one adopted in the permanent census on enterprises, was based on a random forests approach.

4. Exploiting non traditional sources of data in the pandemic

As discussed, the emergency caused in Italy by the COVID-19 pandemic and the measures adopted by the government to contain and mitigate the spread of the virus throughout the country ("lockdown" and "social distancing") had indeed an impact on the continuity and quality of official statistical production of Istat. In response to this situation and in relation to the use of new sources, Istat is intervening on two fronts:

1) Use of alternative sources, methods and techniques to traditional ones to support production processes altered or damaged by the COVID-19 emergency.
2) Development of new official statistics and / or experimental statistics dedicated to understanding the impact of the COVID-19 pandemic.

In relation to 1), during the COVID-19 emergency, Istat increased the frequency of production of the Social Mood on Economy Index (https://www.istat.it/en/archive/219600), an index providing daily measures of the Italian sentiment on the economy, measures that are derived from samples of public tweets in Italian captured in real time. In particular, the index passed from a quarterly to a monthly frequency and was included in the monthly notes on the trend of the Italian economy for March and April 2020. The index was especially useful due to the temporary suspension of the Consumer Confidence survey due to the pandemic. In addition, also with respect to 1), the use of AIS (Automatic Identification System) data was particularly important for the survey on the maritime transport of goods and passengers. In March and April 2020, the lockdown blocked part of the passenger traffic, the restrictions however did not include the transport of goods. AIS helped to understand if the reduction in observations was due to an actual decrease in the phenomenon or to a difficulty in recording by the traditional actors involved. AIS data confirmed the reductions observed through traditional channels and proved to be an important auxiliary source. A relevant role, again with respect to 1), was played by the use of scanner data and web scraped data for consumer prices statistics. The regular production of these statistics was indeed possible by relying on these Big Data channels, underlining the importance of a multi-source and multi-mode approach for the consumer price survey.

With respect to 2) Istat is planning to produce some new statistics for a better understanding of the COVID-19 impact that make use of Big Data sources. There are several projects in this direction at different stages of development, including the use of Mobile Network Operator (MNO) – based statistics to perform pre- and post-COVID comparisons on mobility and tourism statistics, and Twitter-based analyses on post-COVID restart mood, change of topics pre-and post-COVID, etc.
It is worth citing among these projects, a product that Istat developed within the 2021 EU Big Data Hackathon (https://ec.europa.eu/eurostat/cros/BD_Hackathon2021_en) and that won the competition. The product makes use of several datasets, including Google Mobility data, and shows how the pandemic influenced international trade networks in terms of products, countries, means of transport used, and how the policies adopted by the institutions to overcome the crisis have impacted on international trade. Istat is continuing investing on the development of the product that will be proposed as an Istat's experimental statistics in the near future.

5. Conclusions and lesson learned

Due to the COVID-19 pandemic, the Italian NSI had to face unexpected issues in order to maintain the current production of Official Statistics of an adequate quality level, with the additional charge of measuring the impact of the pandemic on households and businesses. The exceptionally high non-response rates reached during the COVID-19 pandemic, mainly due to the impossibility to carry out face-to-face interviews, has required the development of new methodological and data collection innovations to both prevent and treat missing information, including exploring a wider use of non traditional data sources (Big Data). For instance, sampling and estimation strategies have been revised in all situations where data collection moved from face-to-face to (generally mixed) telephone and web modes. Appropriate solutions (e.g. responsive adaptive sampling) were studied to reduce potential bias due to auto-selection effects. The use of different types of Big Data were experimented, either to complement current survey data (to compensate for high non response rates) or to produce new statistics. Eurostat guidelines were taken into account in a number of methodological areas such as time series seasonal adjustment and treatment of missing data in surveys. In addition, during the pandemic Istat has carried out new surveys in order to measure the pandemic's impact on both households and enterprises. In addition, the first Italian epidemiologic survey has been designed and carried out, representing a possible launch of a new area of official statistical production of the Institute.

In terms of lesson learned, the COVID-19 pandemic has represented an important opportunity to understand that Official Statistics have to play a key role also in the management of emergency events, exploiting strengths, competencies and resources to support policy makers in all phases of this type of catastrophic events (i.e. in the response, recovery and prevention phases). In this view, it is now important for Istat to define a strategy to be ready in the case of further crisis periods, planning the necessary investments for structural changes on methodological and data collection surveys approaches, exploiting as much as possible the experiences done so far, and the innovations introduced in statistical processes during the pandemic, often in the form of experimental and temporary solutions. Concerning non traditional data sources, it is fundamental for Istat to further invest on integrating them permanently in current production processes, facing the challenging issues posed by their statistical use, e.g. privacy issues, quality assessment, methodological solutions for treating them and delivering new statistical products based on them. To this aim, organizational changes have been started, with the creation of a new infrastructure (named Centre for Trusted Smart Statistics) with the aim to orientate the strategic investments and priorities on the use of Big Data sources while establishing new collaboration rules inside Istat, with data providers, with other public and private bodies. In this field, particularly important will be also the collaboration within the international statistical community, and continue to invest in new skills and professional figures, like data scientists.

References:

1. Ballin M., Guardabascio B. (2020) Indicatori di monitoraggio FAS. Istat, *internal note* (in Italian).
2. Alleva G., Arbia G., Falorsi P.D., Nardelli V., Zuliani A. (2020). A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design, to be published in *Journal of Official Statistics*, arXiv preprint arXiv:2004.06068.
3. Eurostat (2020) Guidance on treatment of covid-19-crisis effects on data. Methodological note, EUROSTAT, Luxemburg, March 26.
4. Schouten J.G., Cobben F., Bethlehem J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, **35** (1), 101 – 113.
5. Schouten B., Shlomo N. (2017). Selecting Adaptive Survey Design Strata with Partial R-indicators. *International Statistical Review*, **85**, 143-163.