

Technical Note on the methods of the data collection phase¹

Annex to the proposal:

A sample design for reliable estimates of the SARS-CoV-2 epidemic's parameters. Calling for a protocol using panel data²

The document describes the methods of data collection through a sample survey for a reliable estimate of the fundamental parameters of the SARS-CoV-2 epidemic.

Interest Group A

Periodicity of the survey. Appropriate frequency (possibly two weeks) for a period of time related to the evolution of the epidemic.

Territory. To be defined: entire national territory and geographical areas; individual regions; metropolitan cities and provinces; contagion areas, others.

Statistical units subject to the survey. Individuals who came into contact in the previous 14 days with a sample of infected persons previously identified, defined as "contacts". The infected sample is selected from among the people for whom the disease has been diagnosed in the last 48 hours.

Objective of the survey. Carrying out the swab (or other exam) to a sub-sample of contacts.

Detection strategy. 1) Identification of a sample of hospital structures³; 2) Identification and selection of a sample of infected people at them; 3) Identification of the "contacts" of the selected infected people; 4) Detection (if any) at all the "contacts" identified to reach a sub-sample of people on whom to perform the swab (or other examination); 5) Carrying out the swab (or other examination) on the sub-sample.

- Phase 1) Hospitals can be identified in agreement with the Health Authorities on the basis of some criteria. The number of hospitals to be involved could be 6-12 at regional level, or even lower for smaller regions / areas;

¹ Extract from a document edited by Pierluigi Ascani, President of Format Research srl, former professor of Social Research Methodology at Sapienza University of Rome.

² Giorgio Alleva, Full Professor of Statistics, Sapienza University of Rome, Former President of the Italian National Statistical Institute (Istat); Giuseppe Arbia, Full professor of Economic Statistics, Catholic University of the Sacred Heart, Milan, Spatial Data Analysis Expert; Piero Demetrio Falorsi, Former Director of the Methodological Directorate of the Italian National Statistical Institute (Istat), Expert in Sample Designs; Guido Pellegrini, Full Professor of Economic Statistics, Sapienza University of Rome, Expert in Public Policy Evaluation and President of the Statistical Information Guarantee Commission; Alberto Zuliani, Emeritus Professor of Statistics, Former President of the Italian National Statistical Institute (Istat).

³ If you had a list of the infected at national or regional level it would not be necessary to introduce these first stage units.

- Phase 2) A sample of 200 infected people is selected at the hospital facilities, according to criteria to be agreed with the Health Authorities;

- Phase 3) The detection of the infected people is carried out with the aim of identifying all the individuals with whom they came into contact in the 14 days preceding the onset of the disease⁴. The operators in charge of data collection in the field will be faced with infected patients hospitalized and infected in forced quarantine at their homes based on the indication of the hospital structures. Among the former, a distinction must be made between people who are able / unable to hold an interview to find out the names and references of the individuals with whom they have come in contact in the last 14 days. The interviews are carried out face to face in CAPI⁵ mode by subjects equipped with the necessary protective devices. The collected data is transmitted in real time to a centralized database. Those infected who are unable to take the interview are replaced by relatives or friends, identified in collaboration with the hospital structures. Providing an average number of 25 contacts per person infected, the output of the activity consists of a database of approximately 5,000 names.

- Phase 4) Interview with the 5,000 "contacts" identified in phase 3 with the aim of identifying a sub-sample of individuals on whom to perform the swab (or other exam⁶) through a pre-screening questionnaire, proposed in CATI⁷ mode. The output consists of a database consisting of a sub-sample of 1,000 individuals on which to perform the swab.

- Phase 5) Carrying out the swab (or other exam) to the 1,000 identified individuals.

Working times and some aspects of method. Phase 1 is performed upstream of the sample survey. The activities envisaged in phases 2, 3, 4 and 5 can be carried out over the course of two weeks and even a week, except for phase 5, conditioned by the response times of the analysis laboratories. Replications may last a shorter time.

Summary table (referred to the space-time domain of interest)

Hospitals involved	6-12
Infected persons involved	200
Average contacts for each infected person	25
Contacts of the infected people to be interviewed	5.000
<i>(possibly reduced to)</i>	<i>1.000</i>
20% contact sub-sample on which to make the swab (or other exam)	1.000

⁴ The date from which to start the 14-day count must be defined: the day on which the first symptoms were felt, the day the disease was diagnosed, the day of hospitalization.

⁵ CAPI: Computer Assisted Personal Interview.

⁶ This phase could be skipped by randomly selecting the sub-sample of 1,000 units, in addition to the reserves, from the 5,000 names identified in the previous phase.

⁷ CATI: Computer Assisted Telephone Interview.

Interest Group B

Type of investigation. Samples, based on a panel of individuals regularly detected at specific time intervals over a period of time related to the evolution of the epidemic.

Periodicity of the survey. Survey carried out every fourteen days⁸.

Territory. To be identified: entire national territory; single region; contagion area etc.

Statistical units subject to the survey. Presumably healthy individuals, not involved in any way with individuals from interest group A.

Sample size. 1,000 individuals.

Objective of the survey. Carrying out the swab (or other exam) at the panel with the aim of identifying the possible onset of the disease in some individuals within the following 14 days.

Data collection strategy. 1) Construction of the sampling plan and survey questionnaire; 2) Construction of the panel by means of a CATI survey; 3) Carrying out the swab (or other exam) to the individuals of the panel; 4) Identification of the "contacts" in the previous 14 days of the people who were infected; 5) Making the swab to a sub-sample of these "contacts".

- Phase 1) The survey questionnaire must take into account some characteristics of people related to the risk of contagion and suffer serious consequences; among these, in addition to the age and type of family unit, the propensity to travel, in particular by dividing between: a) people who continue to move frequently and therefore more exposed to infection; b) people who have few contacts with others living a regime similar to that of people in quarantine.

- Phase 2) The panel is constructed through a CATI survey, remembering that the 1,000 participants must be available to perform the swab (or other exam) every 14 days (or less) for a period of time related to the evolution of the epidemic⁹. If the survey was mandatory, and even in the case of a justified interest on the part of the population to undergo the test, the panel construction procedure could be quite rapid. If the 1,000 cases were found in 100 different municipalities, twenty teams of two people each - suitably trained by the Civil Protection - would be able to administer the 1000 swabs in a relatively short time, to which would be added that of the laboratory. It is also necessary to provide for the replacement between 20 and 25% of the participants in the panel for routine maintenance¹⁰.

- Phase 3) Carrying out the swab (or other exam) for the members of the panel.

⁸ Possibly more closely, taking into account the median latency interval for the appearance of symptoms after infection.

⁹ In the event that the population should instead be resistant, experience suggests contacting a sample ten times greater than that of the panel, in particular if you want to avoid, for reasons of speed, to replicate contact attempts beyond measure and docking. On the other hand, limiting these attempts can induce a systematic distortion of some characteristics of the panel. There is therefore a trade-off, in the specific context, between speed and representativeness.

¹⁰ To ensure the replacements of the statistical units that decide to abandon the investigation, which are no longer eligible because they are infected, etc.

- Phases 4) and 5) Identification of the "contacts" (in the previous 14 days) of the individuals referred to in phase 3 who may be infected and making the swab on a sub-sample of these "contacts". Presumably, 40 full-blown infected could be detected for each panel detection and then 200 additional swabs could be made to a sub-sample of their contacts.

Summary table (referred to the space-time domain of interest)

Panel (number of persons) on which to perform the swab (or other exam)	1.000
Number of those who could prove to be infected	40
Contacts of persons who could prove to be infected	1.000
Number of contacts on which to make the swab (or other exam)	200
Total swabs (or other tests) to be made	1.200

Other aspects of method

Structured questionnaire for the reconstruction of the "contacts" of the infected

It is considered unlikely that an individual will be able to reconstruct the contacts he had in the 14 days preceding the survey by means of a questionnaire with open questions. The interview should be guided by a reconstruction of the "social networks" in which the person conducts (led) his days before being infected. We can distinguish: primary networks (relatives, friends, neighbors etc.), secondary networks (work colleagues, food shop clerks, pharmacists etc.), formal networks (doctors, law enforcement, health workers etc.) , informal networks (communities, aggregation groups etc.).

Preparation of a data web analysis service

The service could be useful to integrate the data from the interviews regarding the identification of contacts with those from the web and social networks¹¹.

Need for a control room

In a situation in which the epidemic spreads rapidly, one must proceed with extreme speed and with rigorous respect for the scheduled times. This entails the need to provide a "control room" for the detection and analysis of data, a managerial staff with command, control and communication functions. In fact, even a small delay, for example on the part of hospitals in extracting the sample of the infected people (hospitalized or in forced quarantine at their homes), could result in a more than proportional delay on the work schedule of the subsequent stages and with a progressive accumulation effect.

¹¹ Scott, J., *Social Network Analysis. A Handbook*, London, Sage Publications, 2000; Yang, S., Keller, F. B., Zheng L., *Social Network Analysis: Methods and Examples*, London, Sage Publications, 2016.