A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design with a focus on the Italian health system

Giorgio Alleva[1] Giuseppe Arbia [2] Piero Demetrio Falorsi[3] Alberto Zuliani[4]

Abstract: Given the urgent informational needs connected with the pandemic diffusion of the Covid-19 infection, in this paper we propose a sample design to build up a continuous-time surveillance system. With respect to other observational strategies, our proposal has three important elements of strength and originality: (i) it not only aims at providing a snapshot of the phenomenon in a single moment of time, but it is designed to be a continuous survey, repeated in several waves through time, (ii) the statistical optimality properties of the estimators are formally derived in this paper and (iii) it is rapidly operational as it is required by the emergency connected with diffusion of the virus. The sample design is thought having in mind, in particular, the SAR-CoV-2 diffusion in Italy during the Spring of 2020. However, the proposed sampling design is very general, and we are confident that it could be easily extended to other geographical areas and to possible future epidemic outbreaks.

Some keywords: Covid-19 diffusion; efficiency; epidemic monitoring; health surveillance system; sampling design; unbiasedness.

1. Background and purpose

The worldwide urgent need to control the spread of SARS-CoV-2 requires an accurate evaluation of the sources of data on which the estimation of the epidemic's main parameters can be based. Only in this way will we be able to monitor the evolution of the epidemic over time, supporting timely the decision makers in evaluating the effects of the restrictive measures gradually introduced, and the time for their mitigation and removal. In general, this is the way to produce possible future forecasts of the evolutions

[1] Sapienza University of Rome
[2] Catholic University of the Sacred Heart, Milan
[3] Former Director of Methodology at Istat and International Consultant
[4] Emeritus Professor, Sapienza University of Rome

of the disease which are the essential basis for political choices related to an effective healthcare response. Indeed, while some degree of uncertainty is inherent in any statistical modelling, the level of inaccuracy in monitoring the development of the situation can and must be kept under control.

Until now, however, with only few remarkable exceptions (see Section 2) data have been collected, favoring the examination of cases which display symptoms. This situation is described in statistics as "convenience sampling" in the presence of which no sound probabilistic inference is possible (Hansen et al, 1953) More precisely, while in a formal sample design the choice of observations is suggested by a precise mechanism based on the definition of inclusion probabilities of each unit (and, hence, sound probabilistic inference), on the contrary with a *convenience* collection no probability of inclusion can be calculated thus giving rise to over- under-representativeness of the sample units.
In particular, several studies on Covid19 diffusion have clearly shown (e. g. Aguilar et al., 2020; Chugthai et al, 2020; Li, et al. , 2020; Mizumoto et al., 2020a, 2020b and Yelin et al., 2020) that the available data strongly underestimate the number of infected people in that they are unable to capture, e. g., the asymptomatic cases with an obvious overestimation of the lethality rate[5]. On the other hand, a broad-based data collection of medical swabs carried out on a voluntary basis does not constitute a probabilistic sample either[6]. For instance, the practice of systematically collecting observations in the vicinity of supermarkets leads to an over-inclusion of healthy people in the sample, and to a systematic exclusion of those who (either because they are manifesting symptoms or in any case feel weak) have chosen to stay confined at home.
However, it is of crucial importance for government and health officials and for the population to have a clear understanding of the dynamics of the situation in progress in order to take appropriate measures and to guide their individual behaviors. In such a situation, it is essential to set-up a system of data collection which can grant unbiased estimates and statistically significant comparisons through time and across different geographic areas.
During the epidemic to be empirically relevant, not only any sample design has to be technically specified and the properties of the associated estimators have to be proved formally, but it has also to satisfy the following two conditions:

- it has to be implemented as a surveillance system (or strictly related with the existing one) and repeated in several waves rather than a one-shot survey;
- it has to be immediately operational considering the practical implications of the process of data collection.

The latter point is particularly relevant in that the task may prove challenging especially in a situation where all the health operators are employed full time in the emergency operations related to the care of the more severe cases of infected people.

Rather surprisingly the literature on the subject is still extremely poor. Few contributions have suggested the use of crowdsourced data rather than a sample design along with the officially collected data (Leung and Leung, 2020; Sun et al., 2020) although the risk of erroneous inference based on them has been pointed out by Arbia

[5] Lethality rate is given by the proportion of death cases on the infected
[6] https://www.theguardian.com/world/2020/mar/30/immunity-passports-could-speed-up-return-to-work-after-covid-19

(2020). Our aim is to suggest a sample design whose statistical optimality properties are formally proved, but that is also operational and can be immediately put into action by considering the many practical obstacles that may arise in an emergency. Although we have in mind the Italian situation, we are rather confident that the suggested protocol could be easily extended to other countries.

The rest of the paper is organized as follows. In Section 2 we will present a review of the strategies and experiences in progress in the process of data collection. In Section 3 we will present the basic sampling framework of our suggested design by distinguishing two subsets of the population to be surveyed, namely those in which a state of infection has already been verified and those that were in contact with them (group A) and the healthy persons (group B). In Section 4 we focus on the parameters of interest that we aim at measuring with our suggested design on the two groups and we discuss how to disentangle possible overlaps between them whose presence may undermine the statistical properties of the estimation. In Section 5 we provide a general description of the sampling schemes for the two groups and the various operational options to be realized. In Section 6 we prove the unbiasedness of the estimates and define the expression of the sampling variances. Section 7 is devoted to envisaging an extension of the proposed methodology to subsequent waves of data collection to monitor the phenomenon in different moments of time. Section 8 contains some discussion on the efficiency of the estimators. Finally, Section 9 concludes with some practical indications and future research priorities.

2. The data collection of the epidemic: a review of strategies and experiences currently in progress

In the emergency phase connected with the quick and uncontrolled diffusion of the Covid-19 disease, governments and institutions in charge are fully aware that knowledge and understanding of the dynamics in progress represent the central element for establishing how to intervene and the geographical areas in which the intervention is more urgent.

In reviewing the various approaches followed by the different countries until early April 2020, we can identify four strategies and experiences in progress for the estimation of the phenomenon in the entire population.

a) The first consists of *massive test campaigns*, regardless of the presence of symptoms, carried out without following a formal sampling design and essentially aimed at intervening in the outbreaks of the epidemic to identify subjects who are infected, but with no symptoms or only slight symptoms. This was the strategy of South Korea and Hong Kong, as well of United Arab Emirates, Australia, Iceland, Veneto Region in Italy; nowadays this is also the German strategy[7].

b) The second possible strategy consists in *diagnostic tests* through a *probabilistic sample* according to a planned design for the estimation of the phenomena of interest with predetermined precision levels, aimed at estimating the effective size of the infections, including the asymptomatic population. This is the case of the project by the Helmholtz Center for Research on Infections in Germany, on blood testing for antibodies to the Covid-19 pathogen involving over 100,000 individuals (Hackenbroch, 2020).

[7] Italy, France, Spain and UK are still making tests only in the case of specific symptoms and contacts with infected people.

Similarly, in Romania a random sample of 10,500 people living in Bucharest has been planned to detect the infected persons following the directions of the Matei Bals Institute of Infectious Diseases in Bucharest (Romania-insider.com, 2020). Finally, a random selection of people who do not meet the testing criteria will be observed at two Canberra locations by the Australian Capital Territory (Abc, 2020). All these sample surveys have distinct characteristics from both those that are aimed at taking a photograph of the epidemic at a given time (planning to repeat the observation at a later time) and the continuous panel-type surveys with rotated sample for monitoring the evolution of the pandemic over time which constitutes the proposal of this paper and that was suggested in Germany.

c) The third strategy consists of a *specific massive web-survey* collected on basis on individuals and households that decide to participate on a voluntary basis. Some 60,000 Israelis completed the online daily survey developed by the Weizmann Institute, disclosing personal details such as their age, gender, address, general state of health, isolation status and any symptoms they may be experiencing (Rossman et al, 2020). We observed examples of this strategy in Iceland, Estonia and in other countries. The results allow to compare experiences of contagion and testing for people and households with different socio-economic characteristics;

d) Finally, another possible strategy is that of founding the conclusions on pre-existing *sample surveys,* partially modified in order to collect information on the epidemic. Creating an EU 'Corona Panel', as a standardised European sample tests to uncover the true spread of the coronavirus is, indeed, the proposal of the Centre for European Policy Studies, presented by Daniel Gros (2020). The proposal refers, in particular, to the use of the EU-wide sample of the panel of households which participate in the regular surveys on economic and social conditions, called '*EU statistics on income and living conditions*' (EU-SILC). More specifically, Dewatripont et al. (2020) suggest to implement two tests using the EU-SILC panel: the first aimed at assessing whether the subject is currently infected, and the second to test whether the person has become immune due to previous exposure.

Timeliness is crucial. In this respect, the latter strategy seems to guarantee good results for the European Statistical System (ESS). A quick reflection could be made on the feasibility of inserting additional modules in the survey questionnaire of the quarterly Labour Force Survey (LFS), obviously in accordance with the Data Protection Authorities.

As an alternative, the International Labour Organization (ILO) has reached out to the National Statistical Offices (NSOs) to understand the impacts of COVID-19 on their statistical operations, in particular in the domain of labour statistics (ILO, 2020). ILO recommended all countries to consider what additional information could be useful to capture the relevant aspects of the phenomenon. NSOs should consider if some existing topics are of lower priority, and thus can be temporarily removed from the surveys in order to create space for the new questions.

Many countries are experiencing also combinations of the previous different approaches to collect data on the epidemic as well integrating them with administrative data or other official statistical sources. While sample surveys represent a pillar to make inference to the whole population, planning and building integrated informative systems on the epidemic is certainly the right way for a deeper comprehension of the phenomenon. Finally, we observe that in this framework the new data source (such as

mobile phones, web-scraped data and internet-of-things data) should provide a useful contribution.

3. The basic sample framework

In what follows, we aim to propose an observational protocol for the estimation of the people infected by SARS-CoV-2 (Alleva et al., 2020). Starting from a population where it has been ascertained that individuals are infected (the *verified* cases), the aim is to estimate the population that is infected but has not yet been diagnosed (the *asymptomatic* cases). For the purpose of the proposed procedure, the individuals will be preliminarily classified into two sub-groups of interest which we will refer to as: *group A* and *group B*.

Group A is the sub-group consisting of individuals for which a state of infection has been verified (who could be either hospitalized or in compulsory quarantine) and all the people who had contact with them in the previous days. Below we propose to observe the contacts till 14 days before the infection, being this length in time the internationally accepted maximum incubation time. However, the unbiasedness of the sampling strategy we propose is still valid (even if less efficient) if the contacts are reconstructed for a shorter time period (e. g. 7 days). Therefore, this group contains all individuals who are foreseen to be infected and not only those for whom the infection has already been ascertained. They will represent, therefore, both the *apparent* and *latent* dimensions of the epidemic.

Group B contains both healthy people for which the infection is considered *latent* and those who are still in a phase of incubation where the symptoms can become evident in a future moment of time, in the course of a maximum of 14 days.
The rational for this decomposition is related to the feasibility of the observational scheme which we will propose below. Indeed, the proportion of the infected people in the *Group A* is much larger than the one observed in *Group B.* Moreover, the number of verified infected people is known through the data collected by health public authorities. Thus, focusing the investments in observing the contacts of this group maximizes the number of infected people observed in the sample. Nevertheless, it is necessary to observe the *Group B* so as to produce reliable estimate referred to the whole population, which is mandatory for correctly estimating the rate of infected people or the rate of lethality.

Estimates relative to the two sub-groups may be obtained on the basis of a continuous observation in time and following two distinct methodologies, both based on what is known as *indirect sampling*, (Lavalle, 2007; Kiesl , 2016) the same technique that is commonly used for sampling and estimation of rare and elusive populations (Sudman, 1988; Thompson and Seber, 1996).

4. Specification of the total of infected people and its decomposition

In what follows, let U be the population of interest of size N and denote with k ($k = 1, \dots, N$) a person belonging to it. Let v_k be a dichotomous variable which assumes value 1 if state of infection is verified and value 0 otherwise. Let $U_v = \{k \in U : v_k = 1\}$ be the **subpopulation** of U of those for whom the infection is verified and let $\overline{U}_v = U \backslash U_v$ be the complementary subset.

Let y_k be the value for the person k of a variable y which is equal to 1 if the person is infected and 0 otherwise. If $v_k = 1$ then obviously it is also $y_k = 1$, however if $v_k = 0$, then it is possible that either $y_k = 1$ (an infected person for whom the infection has not yet been verified) or $y_k = 0$ (healthy person).

The target parameter of our survey, Y, is the total of infected people (verified or not), that is:

$$(1) \quad Y = \sum_{k \in U} y_k.$$

Let $l_{k,j}$ be the generic entry of a link matrix (k=1,2,..., N; j=1, 2,...N) which is equal to 1 if the individual k had contacts with individual j in the past 14 days and 0 otherwise, with $l_{k,k} = 1$ by definition. Starting from U_v , it is possible to determine the total of y related to the *group A*, $U_A = \{k \in U_v \cup j : l_{k,j} = 1\}$, which includes the subset U_v and all the contacts of those units. We express this formally as:

$$(2) \quad Y_A = \sum_{k \in U_v} \sum_{j \in U_v} \frac{1}{L_{vj}} \, l_{k,j} y_j,$$

where

$$(3) \quad L_{vj} = \sum_{k \in U_v} l_{k,j}$$

is a quantity introduced in order to control the *multiplicity* of the measurement of the y_j among the different k units in U_v in Equation 2.

On the other hand, starting from $\bar{U}_v$ it is possible to determine the total of y related to the *group B*, $U_B = \{k \in \bar{U}_v \cup j : l_{k,j} = 1\}$, which includes $\bar{U}_v$ and all the contacts of these units:

$$(4) \quad Y_B = \sum_{k \in \bar{U}_v} \sum_{j \in U} \frac{1}{\bar{L}_{vj}} \, l_{k,j} y_j,$$

where, analogously to Equation (3), the quantity:

$$(5) \quad \bar{L}_{vj} = \sum_{k \in \bar{U}_v} l_{k,j}$$

is introduced in order to control for the *multiplicity* of the measurement of the y_j in (4) among the different k units in $\bar{U}_v$.

The set U_A and U_B can obviously overlap. Let us define their intersection as the set $U_A \cap U_B = \{j: l_{k,j} = 1 \cap l_{\ell,j} = 1;\ k \in U_v, \ell \in \bar{U}_v\}$. The total of the y_j in $U_A \cap U_B$ is given by:

$$(6)\quad Y_{AB} = \sum_{j \in U_{AB}} y_j$$

$$(7)\quad = \sum_{j \in U_{AB}} y_j \left[\alpha \left(\sum_{k \in U_v} \frac{1}{U_{vj}}\, l_{k,j}\right) + (1-\alpha)\left(\sum_{k \in \bar{U}_v} \frac{1}{\bar{L}_{vj}}\, l_{k,j}\right)\right],$$

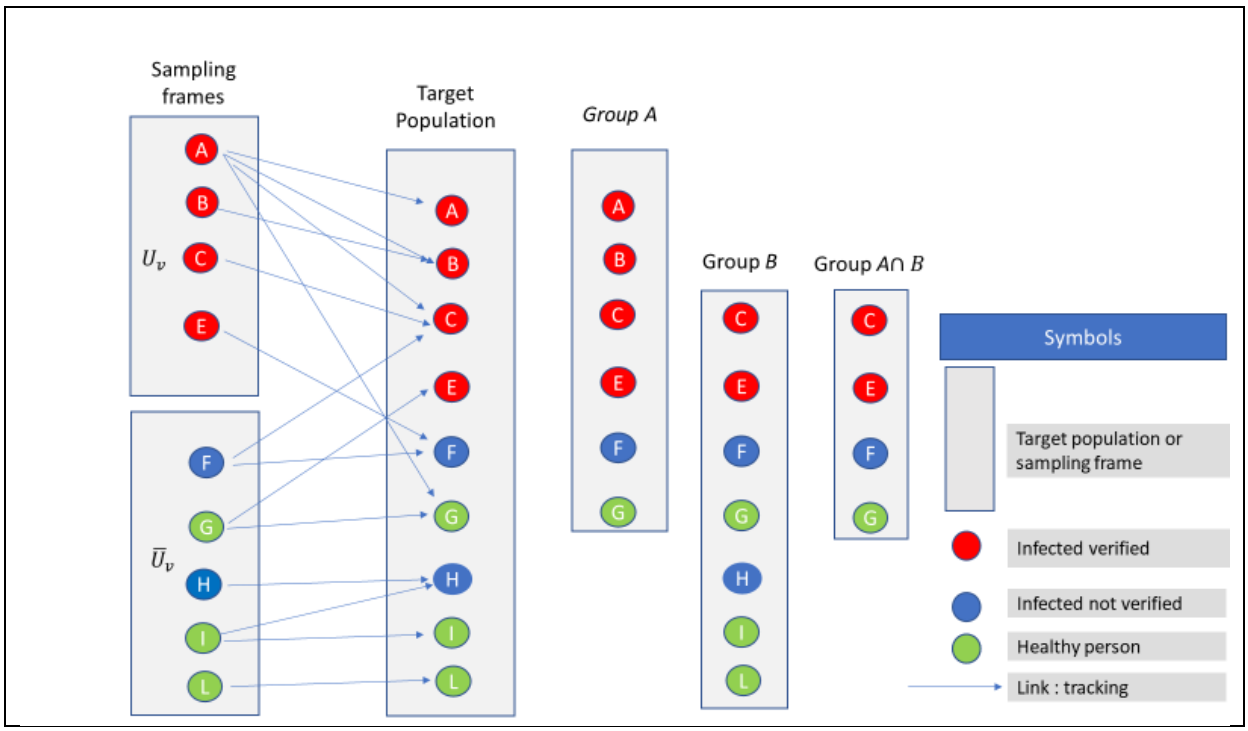
where $0 < \alpha < 1$.

The expression (7) is useful in the phase of integrated estimation which will be illustrated in Section 6 below. Finally, the total Y may then be expressed as

$$(8)\quad Y = Y_A + Y_B - Y_{AB}.$$

The above set-up is illustrated in the Figure 1 below.

Figure 1. Population of interest and its decomposition among the different groups



4. The sampling design

5.1. General description of the sampling schemas

Two independent samples, namely S_v and $\bar{S}_v$, are selected by the two population subsets, U_v and $\bar{U}_v$, which represent the sampling frames. The contacts of infected people in each sample are tracked. The first sample S_v is used for producing an unbiased estimate of Y_A, while $\bar{S}_v$ is used for estimating the total Y_B. The total Y_{AB} is estimated from the both samples.

5.2. Sampling from U_v

The sub-set of the verified infected increases over time. It is therefore necessary to set-up a sampling mechanism which is realized continuously over time. In order to simplify the sampling description, let us suppose that U_v represents the set of the verified infected people in a given time period. The sampling from U_v is carried out in the following phases:

a) a sample S_v is selected without replacement from U_v with inclusion probabilities π_{vk} $(k = 1,2 \dots, \#U_v)$.

b) All the contacts $U_{vk} = \{j \in U: l_{k,j} = 1 \cap k \in S_v\}$ of the individual k selected in S_v are tracked going back 14 days.

c) A sample S_{vk} is selected from U_{vk} without replacement and with equal probabilities of inclusion $\pi_{2v|k}$. We use 2 in $\pi_{2v|k}$ for indicating that this is the inclusion probability of the second stage of the sampling, given the selection of unit k in the first stage.

At the end of the above process, the sample $S_A = S_v \bigcup_{k=1}^{\#S_v} S_{vk}$ is formed with a sampling indirect mechanism including people from both S_v (verified infected people) and $\bigcup_{k=1}^{\#S_v} S_{vk}$ (tracked contacts going back 14 days).

The test to verify the infection is carried out on all the tracked contacts, $\bigcup_{k=1}^{\#S_v} S_{vk}$. Thus, the value of y is known for all the people in S_A.

Remark 1. The phase of tracking all the contacts of a person could be complex and cumbersome. Different solutions are possible. One possibility is to leverage from digital apps allowing epidemic control with digital contact tracing as suggested by Ferretti et al. (2020). Similarly, Ascani (2020) suggests a method based on personal interview. In this case the interviewees must be guided in remembering their contacts by means of a specific structure based on the reconstruction of the "social networks" contacted in the days preceding the infection.

Remark 2. It is clear that for health and wellbeing reasons and to prevent the spread of the infection, it would be best to examine all infected people. However, from the statistical point of view, to obtain estimates of high quality regarding the number of infected persons, this is not strictly necessary. From this point of view, it is more important to concentrate the effort in repeating the examination regularly in time. This effort would be unsustainable with a complete study on the whole population.

5.2.1. Definition of the sampling design

The sampling mechanism for selecting S_v depends on how the data frames for U_v are organized. There are two main possibilities:

Option 1. The data of U_v are available in a centralized data set which can be used for the sample selection,

Option 2. The data of U_v are available only at a decentralized level, so that each healthcare institution has its own list.

The two available options will be discussed in turn in the next two sub-sections.

5.2.1.1 Sampling mechanism for Option 1

If the sampling frame of the infected people is centralized in a unified dataset, one could define a *one stage* sampling design selecting directly the sample units from it. The sampling selection can be carried out with the cube algorithm (Deville and Tillé, 2004, 2005), thus ensuring that the Narain, Hortvitz-Tompson estimates (Narain, 1951; Horvitz and Thompson, 1952) of the selected sample reproduce the known totals of some auxiliary variables (e.g. distribution by sex and age, employment status, geographical distribution etc.This can be expressed as follows:

$$(9) \quad \sum_{k \in S_v} \frac{\mathbf{x}_k}{\pi_{vk}} = \sum_{k \in U_v} \mathbf{x}_k,$$

where $\mathbf{x}_k$ is a vector of P auxiliary variables available for the unit k.

The definition of the optimal inclusion probabilities π_{vk} for the indirect sampling which minimize the cost ensuring a pre-defined level of accuracy for the sampling estimates (or, inversely, minimizing the sampling variances for a given budget) can be determined as illustrated by Falorsi and Righi (2019). Tillé and Wilhelm (2017) suggest to select the sample satisfying Equation (9) through a balanced spatial sampling algorithm which is somehow optimal in maximizing the entropy and minimizing the spatial correlation between neighbouring units (Arbia, 1994; Arbia and Lafratta, 1997, 2002).

Falorsi and Righi (2015) demonstrate that the balancing equations (9) are quite general and allow the definition of a wide class of sampling designs which includes, among the others the Simple Random Sampling Without Replacement (SRSWOR), the Stratified Random Sampling Without Replacement (STSRSWOR), the Stratified random sample with probability proportional to size (PPS), the sample designs with incomplete stratification (SDIS) and many others.

Assuming an *SRS* design, in order to obtain statistical estimates of the number of infected persons in a given *spatial* (the whole national territory or specific geographic area such as, for example, a region) and *temporal* domain (week/day), it would be sufficient to select about 1,000 individuals to test among the contacts of the infected set of persons. This sample size would ensure a reliable estimate with a sampling error around 5% under the assumption that the proportion of infected people in the target population is roughly around 25%.

5.2.1.2. Sampling mechanism for Option 2

If the sampling frames for U_v are available only at healthcare institution level, the selection of units in S_v can be carried out with a two-stage mechanism:

1. **First stage.** A sample S_{1v} of health care institutions is selected from the population of health care institutions (call it U_{1v}). The first stage sample is selected without replacement and with *Probability Proportional to Size*, where the healthcare institution *i* is selected with inclusion probability given by:

$$(10) \quad \pi_{1i} = m\frac{M_i}{M},$$

in which m is the selected number of healthcare institutions to be included in the first stage sampling, M_i is a measure of size of unit *i* and *M* is the overall measure of size. We may define the measure of the size according to different criteria. A good option would be the number of beds available for SARS-CoV-2 patients. The sampling selection of the health care institutions can be carried out with the already quoted "cube algorithm", thus ensuring that the Narain Hortvitz-Tompson estimates of the selected first stage sample reproduce the known characteristics of some auxiliary variables available for the Population U_{1v} (e. g. geographical distribution, number of beds available for SARS-CoV-2 patients etc.). This can be expressed as:

$$(11) \quad \sum_{i\in S_{1v}} \frac{\mathbf{x}_{1v}}{\pi_{1v}} = \sum_{k\in U_{1v}} \mathbf{x}_k,$$

where $\mathbf{x}_{1v}$ is a vector of auxiliary variables for the unit *k.* As suggested for option 1 the sample could be selected, respecting the equation (11), with a balanced spatial sampling algorithm which is optimal, maximizing the entropy and minimizing the spatial correlation of the neighbouring units. Even in this case, the balancing equations (11) allow to define the general class of sampling designs described in Falorsi and Righi (2015).

2. **Second stage.** A fixed number, say $\bar{n}$, of infected people is selected in the sampled institution *drawing the unit* without replacement with a simple random sampling procedure.

In such a way, the sampling is *self-weighting* (Murthy and Sethy, 1965) in the sense that all the units in U_v have an equal probability to be selected. Indeed, the final inclusion probability of the person k to be selected in the healthcare institution *i* is given by the following expression:

$$(12) \quad \pi_{vk} = m\frac{M_i}{M}\frac{\bar{n}}{M_i} = m\frac{\bar{n}}{M}.$$

The *self-weighting* property defines a sampling design which is somehow optimal (Kish, 1966) in the sense that it avoids the negative impact on the sampling variances due to the variability of the sampling weights.

The sampling selection criterion could be based on a time mechanism, which is feasible and, at the same time, easily implementable at a decentralized level. For instance, a sample of infected people could be selected considering those who had access to the healthcare institution within a window of a two-hour time period.

5.3. Sampling from $\bar{U}_v$

In order to estimate the total Y_B an independent *panel* of individuals is selected and monitored repeatedly in time. The operational aspects to be carried out are:

a) First of all, a sample $\bar{S}_v$ is selected without replacement from $\bar{U}_v$ with inclusion probabilities $\bar{\pi}_{vk}$ $(k = 1,2 \dots, \#\bar{U}_v)$.

b) The people in the panel make to a diagnostic test on a regular basis (for example, once a month). If the member k of the panel receives a positive test result (i. e. $y_k = 1$), all their contacts $\bar{U}_{vk} = \{k, j: k \in \bar{S}_v \cap y_k = 1;\ j: l_{k,j} = 1\}$ are tracked, going 14 days back in time.

c) A sample $\bar{S}_{vk}$ is selected from $\bar{U}_{vk}$ without replacemet and with equal inclusion probability $\bar{\pi}_{2v|k}$. We adopted for the second stage inclusion probability, $\bar{\pi}_{2v|k}$, the same notation used for $\pi_{2v|k}$. At end of the whole process, the sample $\bar{S}_B = \bar{S}_v \bigcup_{k=1}^{\#\bar{S}_v} \bar{S}_{vk}$ is formed with an indirect sampling mechanism including people from both $\bar{S}_v$ (people for which the infection status is not known) and $\bigcup_{k=1}^{\#\bar{S}_v} \bar{S}_{vk}$ (tracked contacts going back 14 days of the infected people in $\bar{S}_v$).

5.3.1. A note on some practicalities of the sampling design

The sampling design of the panel could be carried out according to different schemas, depending on the availability of the frame and on other organizational aspects. One possibility is that to form a sub-sample of a regular survey on households carried out by official statistics. Here we assume that the frame on U is represented by a register which is available at a central level, and that for each sample unit we avail a set of auxiliary variables. We assume, furthermore, that in this register the subset $\bar{U}_v$ could also be identified.

In this informative contexts, one stage sampling design could be carried out with optimal inclusion probabilities $\bar{\pi}_{vk}$ determined following Falorsi and Righi (2015, 2019). The sample selection could then be carried out with a balanced spatial sampling algorithm (Tillé and Wilhelm, 2017) ensuring the respect of the following balancing equations:

$$(13) \quad \sum\nolimits_{k \in \bar{S}_v} \frac{\mathbf{x}_k}{\bar{\pi}_{vk}} = \sum\nolimits_{k \in \bar{U}_v} \mathbf{x}_k.$$

where the meaning of the symbols has already been introduced.

Remark 3. The *panel* could be constructed using a two-phase design so as to be able to select by using *pre-screening,* two sub-groups, namely:

- a number of individuals who continue to travel (and are therefore more subject to being infected
- a number of individuals with few contacts who fully observe the prescribed quarantine recommendations.

Remark 4. The two-phase mechanism could be useful if the identification of $\overline{U}_v$ could not carried out. This could be realized in the two-pre-screening phase.

The number of persons involved in the *panel* may be about 1,000 (to obtain around 1,200 tested individuals) for a given *territorial* and *temporal* sampling domain, thus guaranteeing a reliable estimation with a sampling error roughly around 10% assuming that the proportion of infected people in this target population is around 10%.

5.4. Final comments on the sampling design

Comments and good suggestions in this respect came from a discussion with the Portuguese National Statistical Office (INE) a in particular from Francisco Lima, President, Pedro Campos, Director of the Methodology Department and João Lopes from the same dep., with some contributions from Portuguese academia.
Given the complexity of the epidemiology of Covid-19, it may be useful to consider a higher number of sub-groups in Group B. This may become useful in the need of considering heterogeneous models (i. e. considering heterogeneous populations) as it seems to be required for the infectious agent. In particular, it may be of great importance to consider breaking down certain epidemiological parameters into different sub-groups (e. g. transmission coefficient, time to become infectious, proportion of detected cases, time to detection, time to recover). Therefore we suggest to define 4 subgroups considering both the binary factor low-risk/high-risk and the binary factor low-mobility/high-mobility. These are the following:

- a number of individuals not belonging to high-risk groups who continue to travel/work (and are therefore more subject to being infected and infectious);
- a number of people not belonging to high-risk groups with few contacts who fully observe the prescribed quarantine recommendations;
- a number of individuals belonging to high-risk groups who continue to travel/work (and are therefore more subject to being infected and infectious), such as health-care workers;
- a number of people belonging to high-risk groups with few contacts who fully observe the prescribed quarantine recommendations.

As for Group *A*, there might be some advantage in consider the same 4 sub-groups, since the transmission coefficient of each of these sub-groups can be significantly different.
Considering 4 sub-groups in both Group A and B may impact on the sample size which is required to obtain a given sampling error at the sub-group level. The Group B has the potential of studying in detail some crucial "invisible" parameters of the epidemiology of Covid-19 (e. g. proportion of asymptomatic cases, time for symptomatic and asymptomatic to become infectious, and even the proportion of undetected symptomatic cases) and for each of the 4 subgroups independently. Its sample size should be defined with this in mind.

5. Sample estimation of the total of infected people

In each moment of time and in each territorial unit, a direct estimation of the total Y could be obtained as:

$$(14)\quad \hat{Y} = \hat{Y}_A + \hat{Y}_B - \hat{Y}_{AB},$$
$$(14a)\quad = \hat{Y}_A + \hat{Y}_B - \alpha\hat{Y}_{AB}^{A} - (1-\alpha)\hat{Y}_{AB}^{B},$$

where the single elements on the right hand side of (14) are the Generalised Weight Share Method (GWSM) estimates of the corresponding term in (8), $0 \leq \alpha \leq 1$, $\hat{Y}_{AB}^{A}$ and $\hat{Y}_{AB}^{B}$ are the GWSM estimates of the total Y_{AB} obtained by the samples S_A and S_B.

6.1. Estimation of the component $\hat{Y}_A$

The GWSM estimator of the total number of infected people in group *A*, as expressed in Equation (2), is given by

$$(15)\quad \hat{Y}_A = \sum_{k\in S_v} \frac{1}{\pi_{vk}} \sum_{j\in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}}\, l_{k,j} y_j$$
$$= \sum_{k\in S_v} \frac{1}{\pi_{vk}} \hat{Z}_{vk},$$

where:

$$(16)\quad \hat{Z}_{vk} = \sum_{j\in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}}\, l_{k,j} y_j$$

represents the second stage estimate of

$$(17)\quad Z_{vk} = \sum_{j\in U_{vk}} \frac{1}{L_{vj}}\, l_{k,j} y_j.$$

Remark 5. the term L_{vj} in the previous equation corresponds to the total number of contacts of the unit *j* with the verified infected people. It can be collected either with digital contact tracing (Ferretti, 2020) or by the interview.

Proof of the unbiasedness of $\hat{Y}_A$

Denoting with $E(\cdot)$ the operator of sampling expectation, we have

$$(18)\quad E\left(\hat{Y}_A\right) = E\left[\sum_{k\in U_v} \sum_{j\in U_v} \frac{\delta_{vk}}{\pi_{vk}\pi_{2v|k}} \frac{\delta_{2vj|k}}{L_{vj}}\, l_{k,j} y_j\right],$$

where: δ_{vk} is a dichotomous variable being $\delta_{vk} = 1$, if $k \in S_v$ and δ_{vk} =0, otherwise; and $\delta_{2vj|k}$ is a second dichotomous variable being $\delta_{2vj|k} = 1$, if $j \in S_{vk}$ and 0, otherwise.

From Equation (18) we obtain:

$$(19)\quad E(\hat{Y}_A) = \sum_{k\in U_v} \sum_{j\in U_v} \frac{E(\delta_{vk}\delta_{2vj|k})}{\pi_{vk}\pi_{2v|k}} \frac{1}{L_{vj}}\, l_{k,j} y_j.$$

However, since:

$$(20)\quad E(\delta_{vk}\delta_{2vj|k}) = E\left[\delta_{vk} E\left(\delta_{2vj|k} | \delta_{vk} = 1\right)\right] = E\left[\delta_{vk}\pi_{2v|k}\right] = \pi_{vk}\pi_{2v|k},$$

plugging (19) into equation (20), we finally have:

$$E(\hat{Y}_A) = \sum_{k\in U_v} \sum_{j\in U_v} \frac{1}{L_{vj}}\, l_{k,j} y_j = Y_A. \qquad \text{Q. E. D.}$$

Variance of $\hat{Y}_A$

On the basis of the theorem on two stage sampling (Cochran, 1977) the variance of $\hat{Y}_A$ can be expressed as follows:

$$(21)\quad V(\hat{Y}_A) = V_1\left(\sum_{k\in S_v} \frac{1}{\pi_{vk}} Z_{vk}\right) + \sum_{k\in U_v} \frac{1}{\pi_{vk}} V_2\left(\sum_{j\in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}}\, l_{k,j} y_j,\right).$$

In the previous expression the variance is decomposed into the sum of the first stage variance and the first stage expectation of the second stage variance. All the elements of the previous expression can be estimated with standard techniques.

6.2. Estimation of the component $\hat{Y}_B$

The GWSM estimator of the component $\hat{Y}_B$ is given by:

$$(22)\ \hat{Y}_B = \sum_{k\in \bar{S}_v} \frac{1}{\bar{\pi}_{vk}} \sum_{j\in \bar{S}_{vk}} \frac{1}{\bar{\pi}_{2v|k}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j$$
$$= \sum_{k\in \bar{S}_v} \frac{1}{\bar{\pi}_{vk}} \hat{\bar{Z}}_{vk}$$

where the term:

$$(23)\quad \hat{\bar{Z}}_{vk} = \sum_{j\in\bar{S}_{vk}} \frac{1}{\bar{\pi}_{2v|k}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j,$$

represents the second stage estimate of

$$(24)\quad \bar{Z}_{vk} = \sum_{j\in\bar{S}_{vk}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j.$$

Proof of the unbiasedness of $\hat{Y}_B$

To prove the unbiasedness, first of all we have:

$$(25)\quad E\left(\hat{Y}_B\right) = \sum_{k\in\bar{U}_v} \sum_{j\in\bar{U}_{vk}} \frac{E(\bar{\delta}_{vk}\bar{\delta}_{2vj|k})}{\bar{\pi}_{vk}\,\bar{\pi}_{2v|k}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j,$$

where $\bar{\delta}_{vk}$ is a dichotomous variable being $\bar{\delta}_{vk} = 1$, if $k \in \bar{S}_v$ and $\bar{\delta}_{vk}$ =0, otherwise; and $\bar{\delta}_{2vj|k}$ is a dichotomous variable being $\bar{\delta}_{2vj|k} = 1$, if $y_k = 1 \cap j \in \bar{S}_{vk}$ and 0, otherwise.

However, we have:

$$(26)\quad E(\delta_{vk}\delta_{2vj|k}) = E\left[\delta_{vk} E\left(\delta_{2vj|k}|\delta_{vk} = 1\right)\right] = E\left[\delta_{vk}\pi_{2v|k}\right] = \pi_{vk}\pi_{2v|k}.$$

From Equation (25) and (26) it follows:

$$E\left(\hat{Y}_B\right) = \sum_{k\in\bar{U}_v} \sum_{j\in\bar{U}_{vk}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j\,. \qquad \text{Q. E. D.}$$

The term $\bar{L}_{vj}$ corresponds to the total number of contacts of the unit *j* with not verified infected people. Similarly to what happens for the estimation of $\hat{Y}_B$ *t*his information can be collected either with digital contact tracing or by the interview

Variance of $\hat{Y}_B$

The variance may be obtained by simply adapting the expression (22), being:

$$(27)\quad V\left(\hat{Y}_B\right) = V_1\left(\sum_{k\in\bar{S}_v} \frac{1}{\bar{\pi}_{vk}} \bar{Z}_{vk}\right) + \sum_{k\in\bar{U}_v} \frac{1}{\bar{\pi}_{vk}} V_2\left(\sum_{j\in\bar{S}_{vk}} \frac{1}{\bar{\pi}_{2v|k}} \frac{1}{\bar{L}_{vj}}\, l_{k,j} y_j\right).$$

6.3. Estimation of the component $\hat{Y}_{AB}$

From expression (14a), the estimation of the component $\hat{Y}_{AB}$ may be expressed as:

$$(28)\quad \hat{Y}_{AB} = \alpha\hat{Y}_{AB}^{A} - (1-\alpha)\hat{Y}_{AB}^{B},$$

where

$$(29)\quad \hat{Y}_{AB}^{A} = \sum\nolimits_{k\in S_v\cap U_{AB}} \frac{1}{\pi_{vk}} \sum\nolimits_{j\in S_{vk}\cap U_{AB}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}}\; l_{k,j} y_j,$$

$$(30)\quad \hat{Y}_{AB}^{B} = \sum\nolimits_{k\in \bar{S}_v\cap U_{AB}} \frac{1}{\bar{\pi}_{vk}} \sum\nolimits_{j\in \bar{S}_{vk}\cap U_{AB}} \frac{1}{\bar{\pi}_{2v|k}} \frac{1}{\bar{L}_{vj}}\; l_{k,j} y_j.$$

The information on the intersection of the samples with the subpopulation U_{AB} may be collected either during the interview or with digital contact tracing.

Singh and Mecatti (2011) give an in-depth illustration of the different approaches in literature to find the optimal value of α_v in the context of multiple frames surveys. Hartley (1962, 1974) proposed choosing α in (14a) to minimize the variance of $\hat{Y}$. Because the frames are sampled independently, the variance of $\hat{Y}$ is:

$$(31)\quad V(\hat{Y}) = V(\hat{Y}_A) + V(\hat{Y}_B) + \alpha^2\, V(\hat{Y}_{AB}^{A}) + (1-\alpha)^2 V(\hat{Y}_{AB}^{B}) + \\ -2\alpha Cov\,(\hat{Y}_{AB}^{A}, \hat{Y}_A) - 2(1-\alpha)\, Cov\,(\hat{Y}_{AB}^{B}, \hat{Y}_B).$$

Thus, for general survey designs, the variance-minimizing value of α is:

$$(32)\quad \alpha^{opt} = \frac{V(\hat{Y}_B) - Cov\,(\hat{Y}_{AB}^{B}, \hat{Y}_B) + Cov\,(\hat{Y}_{AB}^{A}, \hat{Y}_A)}{V(\hat{Y}_A) + V(\hat{Y}_B)}.$$

Note that if one of the covariances in (32) is large, it is possible for α^{opt} to be smaller than 0 or greater than 1. Hartley (1974) suggests opting for this alternative expression:

$$(33)\quad \alpha^{*} = \frac{V(\hat{Y}_B)}{V(\hat{Y}_A) + V(\hat{Y}_B)}.$$

Unbiasedness and variance. The proof of unbiasedness and the calculation of the variance of the estimator $\hat{Y}_{AB}$ are straightforward extensions of what has been illustrated in sections 6.1 and 6.2.

Remark 7. Lavallé and Rivest (2012) propose to estimate the total Y with the *Generalised Capture-Recapture Estimator* (GCRE), which makes a joint use of capture-recapture *Petersen* estimator with GWSM estimators. In our context, the GCRE estimator may be expressed as:

$$(34) \qquad \hat{Y}_{GCRE} = \frac{\hat{Y}_A \, \hat{Y}_B}{\hat{Y}_{AB(S_A \cap S_B)}},$$

where $\hat{Y}_{AB(S_A \cap S_B)}$is the estimate of Y_{AB} computed on the basis of the units observed in the intersection sample $S_A \cap S_B$ in which the sampling weights for producing the estimates from $S_A \cap S_B$ are given in formula (11) in the above mentioned paper.

7. Sample design for the follow-up of the survey in subsequent waves

The observational scheme proposed in the above sections is set up as a cross sectional survey. However, it can be adapted to monitoring the evolution of the number of infected people over time, according to a mechanism which is updated as in a chain mechanism time after time. While an in-depth study of this aspect deserves a separate study, we limit ourselves to introduce here the topic and to provide some initial indication.

Let consider two consecutive points in time, say $t = 0$ and $t = 1$.

The person k verified as infected at time 0, hence denoted as $v_{0,k} = 1$, may still be infected ($v_{1,k} = y_{1,k} = 1$) or she/he may no longer be infected ($y_{1,k} = 0$) by *death* (denoted by the dichotomous variable $d_{1,k} = 1$) or *healing* (denoted by the dichotomous variable $h_{1,k} = 1$).

The total of the y variable at time 1, may then be defined as:

$$(35) \quad Y_1 = Y_0 + \Delta D_{0\to1} + \Delta H_{0\to1} + \Delta Y_1,$$

where Y_0 is the total number of infected at time 0 and:

$$(36) \quad \Delta D_{0\to1} = \sum_{k\in U} y_{0,k}\, d_{1,k}, \qquad \Delta H_{0\to1} = \sum_{k\in U} y_{0,k}\, h_{1,k}\,, \Delta Y_1 = \sum_{k\in U} (1 - y_{0,k})\, y_{1,k}.$$

In equation (36) the quantity $(Y_0 + \Delta D_{0\to1} + \Delta H_{0\to1})$ indicates the total number of verified infected people at time 0 who are still infected at time 1, while the quantity , ΔY_1 denotes the total number of *new* infected.

The updating of the sampling structures illustrated in the previous sections allows to obtain a direct estimate of each of the components of (35), as illustrated in the Figure 2.

The total ΔY_1can be estimated, as described in Section 5, using two sources of data, namely:

- the sample $S_{v,1}$, which automatically captures the new entrances in the population of the verified infected at time 1, $\Delta U_{1,v}$, since the sampling selection is carried out

continuously over time on this population. Then a sample of their contacts could be carried out as described in section 4.2, obtaining the sample $S_{1,A}$;

- the panel $\bar{S}_v$ which is updated over time, since the tests carried out at time $t = 1$ on the individuals of $\bar{S}_{v,0}$ individuate the *new infected* people of the panel. Then tracking their contacts allows to obtain the sample $S_{1,B}$.

The estimation of the totals $(Y_0 + \Delta D_{0\to1} + \Delta H_{0\to1})$ can be obtained by following up the health status of the infected people captured in the samples $S_{0,A}$ and $S_{0,B}$ of time 0. The estimates are then obtained with the sampling weights computed at time 0.

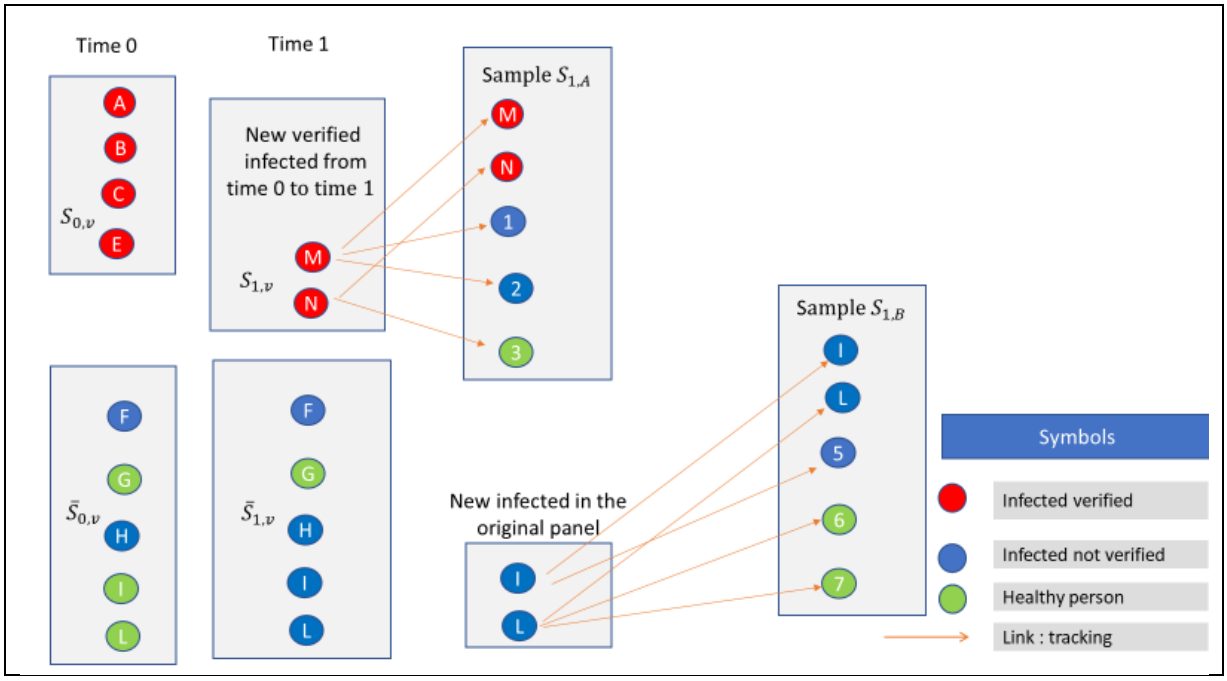
Therefore, we have:

(37) $\hat{Y}_1 = \hat{Y}_0 + \widehat{\Delta D}_{0\to1} + \widehat{\Delta H}_{0\to1} + \widehat{\Delta Y}_1,$

where $\hat{Y}_0, \widehat{\Delta D}_{0\to1}, \widehat{\Delta H}_{0\to1}, \widehat{\Delta Y}_1$ are the direct estimates of the corresponding quantities $Y_0, \Delta D_{0\to1}, \Delta H_{0\to1}, \Delta Y_1$. The above mechanism can be updated in a chain mode, thus obtaining the estimate for the time $t > 1$ as:

(38) $\hat{Y}_t = \hat{Y}_{t-1} + \widehat{\Delta D}_{t-1\to t} + \widehat{\Delta H}_{t-1\to t} + \widehat{\Delta Y}_t.$

Figure. 2. Follow up of samples over time

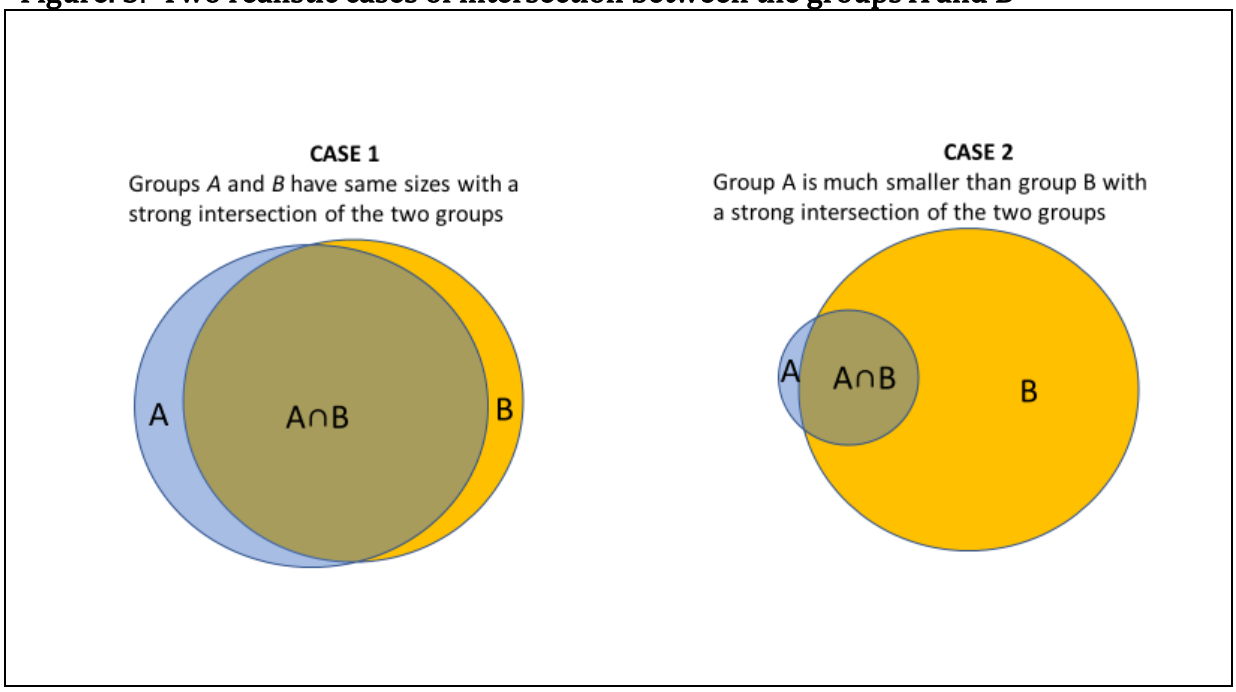


8. A note on the efficiency of the estimators

In order to derive the efficiency of the estimators we need to specify different cases that may occur, related to the intersection of Group A and Group B. Here we consider only two rather realistic cases, which are illustrated in Figure 3.

- **Case 1.** Groups A and B have same sizes with a strong intersection between the two.
- **Case 2.** Group A is much smaller than group B with a strong intersection between the two groups.

Figure. 3. Two realistic cases of intersection between the groups A and B



Case 1

Let us start from the expression (32) of the variance of $\hat{Y}$, the expression of which is reported here for convenience:

$$V(\hat{Y}) = V(\hat{Y}_A)\,(1 + \alpha^2 - 2\alpha) + V(\hat{Y}_B)[1 + (1-\alpha)^2 - 2(1-\alpha)].$$

In the situation described under Case 1, it is possible to introduce the following approximation:

$$(39)\quad Cov\left(\hat{Y}_{AB}^{A}, \hat{Y}_A\right) \cong V\left(\hat{Y}_{AB}^{A}\right) \cong V(\hat{Y}_A);\; Cov\left(\hat{Y}_{AB}^{B}, \hat{Y}_B\right) \cong V\left(\hat{Y}_{AB}^{B}\right) \cong V(\hat{Y}_B).$$

Using the above approximation in (32), we have

$$(40)\quad V(\hat{Y}) = V(\hat{Y}_A)\,(1 + \alpha^2 - 2\alpha) + V(\hat{Y}_B)[1 + (1-\alpha)^2 - 2(1-\alpha)] = \\ = V(\hat{Y}_A)\,(1-\alpha)^2 + V(\hat{Y}_B)\alpha^2.$$

Since $0 < \alpha < 1$, we have a gain in efficiency with respect to the sum of the variances $V(\hat{Y}_A)$ and $V(\hat{Y}_B)$.

Case 2

In this second instance, it is possible to assume the following approximation:

$$(41)\quad Cov\left(\hat{Y}_{AB}^{A}, \hat{Y}_A\right) \cong V\left(\hat{Y}_{AB}^{A}\right) \cong V\left(\hat{Y}_A\right);\ \ Cov\left(\hat{Y}_{AB}^{B}, \hat{Y}_B\right) \cong V\left(\hat{Y}_{AB}^{B}\right) \cong \gamma V\left(\hat{Y}_B\right),$$

where

$$(42)\quad \gamma = \frac{V\left(\hat{Y}_{AB}^{B}\right)}{V\left(\hat{Y}_B\right)}$$

since $\hat{Y}_B \gg \hat{Y}_{AB}^{B}$.

Using the above approximation in (32), we have:

$$V\left(\hat{Y}\right) = V\left(\hat{Y}_A\right)(1 + \alpha^2 - 2\alpha) + V\left(\hat{Y}_B\right)[1 + (1-\alpha)^2\gamma - 2(1-\alpha)\gamma]$$
$$(43)\quad = V\left(\hat{Y}_A\right)(1-\alpha)^2 + V\left(\hat{Y}_B\right)(1 + \alpha^2\gamma - \gamma),$$

being $1 + \alpha^2\gamma - \gamma < 1$, since $\alpha^2 < 1$.

9. Conclusions and future challenges

The aim of this paper is to draw the attention of researchers and decision makers on the need of observing the characteristics of the Covid19 pandemic through a formal sample design thus overcoming the limitations of data collected on a convenience basis. Only in this way will we be able to produce both reliable estimates of the current situation and forecasts of the future evolution of the epidemic so as to take empirically grounded decisions about public health monitoring and surveillance, especially in the phase of the exit from the epidemic peak and of relaxation of the quarantine measures.

In such a situation, it is essential to set up a system of data collection which allows statistically significant comparisons through time and across different geographic areas, by taking into account the different economic, demographic, social, environmental and cultural contexts.

We believe that a clear knowledge of the phenomenon is necessary also for the awareness and behaviour to be adopted by the population. Trust and sharing must be grounded on a solid information base.

In comparison with other possible observational strategies the proposal has three elements of strength, namely:

- the relevance; the proposed sample scheme, designed to capture most of the infected people through an indirect sampling mechanism, not only aims at providing a snapshot of the phenomenon in a single moment of time, but it is designed so as to become a continuous survey, repeated in several waves through time. It contributes to implement a statistical surveillance system on the epidemic to be integrated with the existing systems of surveillance managed by the health authorities;
- the statistical quality of the proposed estimators; in the paper the properties of the estimators have been formally derived so as to guarantee their reliability (unbiasedness and efficiency);
- the timeliness; the sample design is rapidly operational as it is required by the emergency we are experiencing; indeed, the paper represents the statistical formalization of a recent proposal (Alleva et al., 2020) that has been accompanied by a technical note which describes the different phases in which it is divided, the subjects involved and the crucial points for its success (Ascani, 2020).

Although our effort to progress on the subject in this phase of emergency, there is floor for a lot of methodological statistical research for setting up statistical instruments for producing reliable and timely estimates of the phenomenon. Indeed, from a methodological point of view, while in the paper we have fully derived the properties of the estimators in the cross-sectional case, the properties of the estimator in subsequent waves still need to be proved formally. Among other aspects to be developed, we mention those related to time and spatial correlations, useful both for modelling the phenomenon and for designing efficient spatial sampling so as to achieve the same level of precision with fewer sample units (Arbia and Lafratta (2002). A specific extension of the spatial sampling techniques to be further developed is the use of capture/recapture techniques (Borchers, 2009) which would require an overlap of the samples of groups *A* and *B*. Other general aspects to be developed with different specialists are the integration of the statistical system we propose with the health authority's surveillance system for the infected and the use for statistical purposes of the contact-tracking devices, both in the identification of contacts and in the measure of the propensity to travel and of the connected risks.

The sample survey we proposed may represent therefore part of an integrated information system that allows to respond to a plurality of objectives like: (*i*) to monitor the evolution of the phenomenon and to estimate its diffusion patterns; (*ii*) to estimate correctly the actual number of infected people and the rates of asymptomatic cases, recovery and lethality; (*iii*) to assess the effects of the restrictive measures introduced; (*iv*) to orient the mitigation and removal times of the same measures.

In order to be able to build up an integrated statistical system it is necessary to integrate different resources. The surveillance system should merge in a unified database the information collected by the administrative institutions when receiving and treating individuals that have turned to the healthcare system together with the statistical information collected on purpose with the aim to accurately measure the diffusion of the infection and, finally, with those obtained through new sources (such as, for instance

mobile phones, internet-of-things, webscraping, drones images) for tracking the movements of people and of their contacts.

From an operational point of view, the sample design described in detail in Section 5 needs to be accompanied by the rapid definition of some key points:

- a *control room* that ensures the necessary inter-institutional collaboration to guide field operations;
- the medical testing procedure to consider for the selected population (swabs, blood testing and DNA mapping);
- the legal framework to assure the data collection and the analysis of personal data by interviews and digital disposals;
- the geographical-temporal estimate domains of interest and the sample dimension on the basis of the informative needs and the available financial and organizational resources;
- the frequency of sampling for groups A and B, as well as the length of stay in the panel of group *B*;
- the socio-demographic characteristics, living condition and mobility behaviors to be collected at individual and familiar level to shed light on relative risks and to evaluate the effects of the policies adopted to modify the evolution of the epidemic.

This can only be achieved if epidemiologists, virologists, administrators of healthcare institutions work in conjunction with experts in mathematical and statistical modeling and forecasting and in the evaluation of public policies.

We designed the sampling mechanism having in mind the Italian situation. In adopting the suggested strategy, different countries may require adjustments taking into account the peculiarities of the specific health system and institutional framework. In this direction it will be essential the contribution of the National Statistical Offices, as well as common actions and sharing experiences at the European and worldwide level.

The emergency connected with the diffusion of Covid19 is an incredible occasion for building up a solid informative infrastructure for researchers and decision makers. We must also feel the duty and responsibility to prepare to face possible future outbreaks of pandemics in an informed way.

References

Abc (2020). *Random coronavirus testing to begin in Canberra next week at drive-through centre and clinic*. Abc net 03-04-2020. https://www.abc.net.au/news/2020-04-03/random-coronavirus-testing-begins-in-canberra/12119364.

Alleva, G., Arbia, G., Falorsi, P. D., Pellegrini, G., Zuliani, A. (2020). A sample design for reliable estimates of the SARS-CoV-2 epidemic's parameters. Calling for a protocol using panel data. https://web.uniroma1.it/memotef/sites/default/files/Proposal.pdf .

Alleva, G. (2017). The new role of sample surveys in official statistics, ITACOSM 2017, The 5th Italian Conference on Survey Methodology, 14 giugno 2017, Bologna, https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf.

Aguilar, J. B., Faust, J. S. Westafer, L. M. and Gutierrez, J. B. (2020) Investigating the Impact of Asymptomatic Carriers on COVID-19, medXiv, doi: https://doi.org/10.1101/2020.03.18.20037994 .

Arbia, G. (1994) Selection techniques in sampling spatial untis, *Quaderni di statistica e matematica applicata alle scienze economico-sociali*, XVI, 1-2, 81-91.

Arbia, G. (2020) A Note on Early Epidemiological Analysis of Coronavirus Disease 2019 Outbreak using Crowdsourced Data, arXiv:2003.06207 .

Arbia, G. and Lafratta, G. (1997) Evaluating and updating the sample design: the case of the concentration of SO_2 in Padua, *Journal of Agricultural, Biological and Environmental Statistics*, 2, 4, 1997, 451-466, IF 1,235.

Arbia, G. and Lafratta, G. (2002) Spatial sampling designs optimized under anisotropic superpopulation models, *Journal of the Royal Statistical Society series c – Applied Statistics*, 51, 2, 2002, 223-23.

Ascani, P. (2020). Technical Note on the methods of the data collection phase for a proposal of sample design for reliable estimates of the *epidemic's parameters* of SARS-CoV-2. https://web.uniroma1.it/memotef/sites/default/files/TechNote.pdf

Borchers, D. (2009) A non-technical overview of spatially explicit capture–recapture models. *Journasl of Ornitholo*gy, 152**,** 435–444. https://doi.org/10.1007/s10336-010-0583-z

Chughtai, A.A. and Malik, A.A., (2020). Is Coronavirus disease (COVID-19) case fatality ratio underestimated?. *Global Biosecurity*, 1(3).

Cochran, W.G. (1977) *Sampling Techniques*. Wiley. New York.

Deville, J.-C. and Tillé, Y. (2004). Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91, 893-912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Dewatripont, M, M Goldman, E Muraille and J-P Platteau (2020). *Rapidly identifying workers who are immune to COVID-19 and virus-free is a priority for restarting the economy*, VoxEU.org, 23 March.

Falorsi P. D., Righi P. (2015), Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey methodology*, vol. 41. p. 215-236 , ISSN: 0714-0045.

Falorsi P. D., Righi P., Lavallée P. (2019). Optimal Sampling for the Integrated Observation of Different Populations. *Survey methodology*, Vol. 45, No. 3, pp. 485-511. *Statistics Canada*, Catalogue No. 12-001-X.

Ferretti, L, Wymant, C., Michelle Kendall, Lele Zhao, Anel Nurtay1, Lucie Abeler-Dörner, Michael Parker, David Bonsall, Christophe Fraser, (2020), Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, *Science* 31 Mar 2020,DOI:10.1126/science.abb6936https, //science.sciencemag.org/content/early/2020/03/30/science.abb6936 .

Giuliani, D., Dickson, M. M., Espa, G. and Santi, F. (2020) Modelling and predicting the spatio-temporal spread of Coronavirus disease 2019 (COVID-19) in Italy, arXiv:2003.06664 .

Gros, D (2020). "*Creating an EU 'Corona Panel*': Standardised European sample tests to uncover the true spread of the coronavirus" VoxEU.org, 28 March.

Gross, B, Zhiguo Zheng, Shiyan Liu, Xiaoqi Chen, Alon Sela, Jianxin Li, Daqing Li, Shlomo Havlin (2020) Spatio-temporal propagation of COVID-19 pandemics.

Hackenbroch, V (2020). *Große Antikörperstudie soll Immunität der Deutschen gegen Covid-19 feststellen*, Spiegel 26-03-2020 https://www.spiegel.de/wissenschaft/medizin/coronavirus-grosse-antikoerper-studie-soll-immunitaet-der-deutschen-feststellen-a-c8c64a33-5c0f-4630-bd73-48c17c1bad23?d=1585300132&sara_ecid=soci_upd_wbMbjhOSvViISjc8RPU89NcCvtlFcJ .

Hartley, H. O. (1962), Multiple Frame Surveys," *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H. O. (1974), Multiple Frame Methodology and Selected Applications, *Sankhya*, Ser. C, 36, 99.

Horvitz, D.G. and Thompson, D.l. (1952). A generalisation of sampling without replacement from finite-universe. J. Amer. Statist. Assoc., 47,663-685.

International Labour Organization (2020). COVID-19 impact on the collection of labour market statistics https://ilostat.ilo.org/topics/covid-19/covid-19-impact-on-labour-market-statistics/.

Kaiyuan Sun, Jenny Chen and Cecile Viboud, C. (2020) Early epidemiological analysis of coronavirus disease 2019 outbreak using crowdsourced data: a population level observational study, thelancetdigitalhealth, https://doi.org/10.1016/S2589-7500(20)30026-1 .

Kiesl, Hans. (2016). Indirect Sampling: A Review of Theory and Recent Applications. AStA Wirtschafts- und Sozialstatistisches Archiv. 10. 10.1007/s11943-016-0183-3.

Kish, L. (1965). *Survey Sampling*, Wiley. New York.

Lavallée, P., Rivest L.P: (2012). Capture–Recapture Sampling and Indirect Sampling. *Journal of Official Statistics*, Vol. 28, No. 1, 2012, pp. 1–27.

Lavallée, Pierre P (2007) *Indirect Sampling*, springer series in statistics.

Leung, G. and Leung, K (2020) Crowdsourcing data to mitigate epidemics, the lancet digital health, Open Access Published: February 20,2020DOI: https://doi.org/10.1016/S2589-7500(20)30055-8 .

Li, R., Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, Jeffrey Shaman (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2), *Science* 16 Mar 2020, eabb3221, DOI: 10.1126/science.abb3221 .

Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020a). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, *25*(10), 2000180. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180 .

Mizumoto, K., Katsushi, K. Zarebski, A. and Gerardo (2020b) Estimating the Asymptomatic Proportion of 2019 Novel Coronavirus onboard the Princess Cruises Ship, 2020, medRxiv, .https://doi.org/10.1101/2020.02.20.20025866doi.

Murthy M. N. and Sethi V. K. (1965) Self-Weighting Design at Tabulation Stage *Sankhyā: The Indian Journal of Statistics, Series B,* 27, 1-2, 201-210.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. J. Ind. Soc. Agril. Statist., 3,169-174.

Romania-insider.co (2020). *Coronavirus in Romania*: Over 10,000 Bucharest residents will be tested for Covid-19 as part of a study. 03-04-2020. https://www.romania-insider.com/coronavirus-romania-bucharest-testing-streinu-cercel .

Rossman, H., Ayya Keshet, Smadar Shilo, Amir Gavrieli, Tal Bauman, Ori Cohen, Ran Balicer, Benjamin Geiger, Yuval Dor, Eran Segal (2020). *A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys*. https://doi.org/10.1101/2020.03.19.20038844.

Singh, A.C. & Mecatti, F.(2011). Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, 27(4): 633–650.

Sudman, S., Monroe G. Sirken, M. G. and Cowan, C. D. (1988), Sampling Rare and Elusive Populations, Thompson S.K., Seber G.A.F. (1996), *Adaptive Sampling*. *Science*, New Series, 240, 4855 991-996. ISBN: 978-0-471-55871-2 July 1996

Yelin, I. , Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagam Gandali, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, Roy Kishony (2020) Evaluation of COVID-19 RT-qPCR test in multi-sample pools, medRxiv, 27 march, 2020.doi: https://doi.org/10.1101/2020.03.26.20039438.