A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design

Giorgio Alleva[1] Giuseppe Arbia [2] Piero Demetrio Falorsi[3] Vincenzo Nardelli[4] Alberto Zuliani[5]

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, the 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, taking into account different target variables in different stages of the development of the epidemic; (ii) the statistical optimality properties of the proposed estimators are formally derived and tested with a Monte Carlo experiment and (iii) it is rapidly operational as it is required by the emergency connected with the 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, it is very general, and we are confident that it could be easily extended to other geographical areas and to possible future epidemic outbreaks. Formal proofs and a Monte Carlo exercise highlight the estimator is unbiased with a higher efficiency with respect to the simple random sampling scheme.

Some keywords: Covid-19 diffusion; relative 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 the decision makers in evaluating the effects of the restrictive measures gradually introduced and the time for their mitigation and removal.

[1] Sapienza University of Rome, giorgio.alleva@uniroma1.it
[2] Catholic University of the Sacred Heart, Milan
[3] Former Director of Methodology at Istat and International Consultant
[4] Bicocca University, Milan
[5] Emeritus Professor, Sapienza University of Rome

In general, this is the way to produce possible future forecasts of the evolutions of the disease which are the essential basis for 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.

The objective of the proposed method is the definition of an observational protocol for observing the epidemic over time and providing statistically significant estimates of the size of the different components of the population identified concerning the SARS-CoV-2 epidemy. Moreover, we aim to propose a dynamic monitoring tool that can be suitably calibrated both in the growth phase of the infections., and in the decreasing phase, with estimates extended to the parameters of the progressive immunization of the population. All estimates can be produced with associated reliability measures.

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, by sound probabilistic inference), 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[6]. 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[7]. 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 both government and health officials and for the population to have a clear understanding of the dynamics of the epidemics while it is in progress so that the government can take the appropriate measures and guide the 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 to be a one-shot survey;
- it has to be immediately operational considering the practical implications of data collection.

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

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); the risk of erroneous inference based on them has been pointed out by Arbia (2020), Di Gennaro et al. (2020) and Ioannidis (2020). Our aim is to suggest a sample design whose statistical optimality properties are formally proved, that is also operational and can be immediately put into action taking into account 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 present a review of the strategies and experiences in progress in the process of data collection until early April 2020. In Section 3 we 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 who were in contact with them (group A) and the healthy persons (group B). The different role of the two groups in monitoring the infections in different stages of the epidemic is also discussed. In Section 4 we focus on the parameters of interest that we aim at measuring with the suggested sample design on the two groups and we discuss how to disentangle possible overlaps between them whose presence may undermine the statistical properties of the estimations. 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 derive 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 and stages of the epidemic. Section 8 contains some discussion on the efficiency of the estimators. Section 9 illustrates the empirical results of a simulation study. Finally, in Section 10 we suggest some practical indications and future research priorities. The formal proofs are relegated to the Appendix.

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, 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 in which geographical areas the intervention is more urgent.

In reviewing the 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[8]. The big limit of this approach is the impossibility to make statistical inference of the results to the whole population.

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, based on blood testing for antibodies to the Covid-19 pathogen and involving over 100,000 individuals (Hackenbroch, 2020). 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 are cross-sectional, useful to measure the infection in a precise instant. However, they have distinct characteristics from the continuous-panel-type surveys with a rotated sample for monitoring the evolution of the pandemic over time which constitutes the proposal of this paper.[9]

c) The third strategy consists of a *specific massive web-survey* collected 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 the same 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. As for strategy a), the self-selection in the sample makes not possible to extend the results to the whole population.

d) A further possible strategy is to use *pre-existing sample surveys,* partially modified in order to collect information on the epidemics. 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.

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

[9] UK and Italy recently realized sample surveys at national level to estimate the real prevalence rate of the infection (ONS, 2020; Istat, 2020). A critical review on the available data on Covid-19 and on the Italian sample survey project is contained in Alleva and Zuliani 2020, and Alleva, 2020.

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 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 sources (such as mobile phones, web-scraped data and internet-of-things data[10] used bto trace the movements of people) 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 shows no symptoms (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 has been diagnosed, 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 breakdown of the population is related to the feasibility of the observational scheme which we propose. 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

[10] For example, data collection through images, useful for tracking movements of people or vehicles, or to detect gatherings in specific places.

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 and 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 estimation of rare and elusive populations (Sudman, 1988; Thompson and Seber, 1996).

It is important to underline that the distinctive element of our proposal lies in the estimate of the infected population obtained by combining together the results obtained through two samples drawn from the populations A and B, that can establish a different role in relation to the various development phases of the epidemics (in terms of sample size and/or type of diagnostic assessment to be carried out).
At the beginning of the epidemics, the infection has the characteristics of rapidity, unpredictability of the level of spread, apparent concentration in certain geographical areas and categories of subjects. The response of the health system and the containment measures are not yet codified, as well as the responsible behaviour of the population. In this phase, an investigation strategy based on indirect sampling appears to be coherent, starting from the immediate surroundings of subjects that have a confirmed infection. This is the sampling strategy proposed for group A which, in addition to the estimation of a rare phenomenon in the population, provides also an immediate (and continuous over time) response to the epidemics where it explicitly manifests itself.

On the other hand, in order to measure the intensity and the evolution of the phenomenon for larger territorial domains and in general with reference to relevant characteristics of people (gender, age, educational qualification, professional status and more), a traditional population panel survey with sample rotation aimed at group B can be carried out, associated with an indirect sampling mechanism in order to trace and to sample the individuals who came into contact with the infected people founded in this second sample. This panel survey becomes fundamental in the phases which follow the epidemic peak in order to measure not only the effective reduction of infections (and therefore to test the positive effects of the containment measures), but also the proportion of population that had contacts in the past with the virus. In the declining phase of the epidemics (which naturally does not exclude the resumption of infections in specific territories and environments), the role of the sample from the population group B is fundamental and representative of the entire population followed over time. On the other hand, also the diagnostic test must be identified taking into account the different importance that the infected population and the population susceptible to infection assume in the various phases of the epidemics. From an operational point of view, it seems convenient to rely on the nasopharyngeal swab for the sample of contacts in group A, regardless of the phase of the epidemics. For the panel survey, the serological examination may be more convenient, in particular in the declining phase, together with a part of the sample to be foreseen, to which the swab is also administered[11].

[11] It is important to emphasize that, while the swab allows to estimate the infected population at a given moment of time, the serological test allows to estimate the population that had contact with the virus, without a time reference. On the other hand, both diagnostic tools provide answers affected by errors and consequently the estimates must be considered in probabilistic terms. In particular, in order to ensure greater reliability of the results, while for the swab the health protocols require its repetition through time so as to ascertain the healing of those who contracted the virus, for the serological examination, diagnostic kits can be considered that ensure

The combination of the two sampling strategies (with different weights in the ascending and descending phases of the epidemics) represents the competitive advantage of our proposal: a dynamic monitoring tool designed in order to be suitably calibrated both in the growth phase of the infections, providing estimates according to different categories of severity, and in the decrease phase, with estimates extended to the parameters of the progressive immunization of the population.

The advantage over a strategy based exclusively on indirect sampling or only on the panel sample can be measured in terms of greater efficiency (and therefore accuracy of estimates), and lower investigation costs to achieve predetermined levels of precision.

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

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 $U_C = U \backslash U_v$ 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,

$$U_A = \left\{ j \in U: \sum_{k \in U_v} l_{k,j} \geq 1 \right\}$$

which includes the subset U_v and all the contacts of those units. We express this formally as:

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

predetermined levels of specificity and sensitivity. For a discussion of the impact of these errors in epidemic stages characterized by a different base rate of infection, see Fuggetta (2020).

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 U_C it is possible to determine the total of y related to the *group B*,

$$U_B = \left\{ j \in U : \sum_{k \in U_C} l_{k,j} \geq 1 \right\}$$

which includes U_C and all the contacts of the infected people of U_C:

$$(4)\quad Y_B = \sum_{k \in U_C} y_k \sum_{j \in U} \frac{1}{L_{Cj}}\, l_{k,j} y_j,$$

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

$$(5)\quad L_{Cj} = \sum_{k \in U_C} y_k\, l_{k,j}$$

is introduced in order to control for the *multiplicity* of the measurement of the y_j in (4) among the different k units in U_C.

The set U_A and U_B can obviously overlap. Let us define their intersection as the set

$U_{AB} = U_A \cap U_B = \{ j \in U : L_{vj} L_{Cj} \geq 1 \}$.

The total of the y in $U_A \cap U_B$ is given by:

$$(6)\ Y_{AB} = \sum_{j \in U : L_{vj} L_{Cj} \geq 1} y_j$$

We may obtain alternative expressions of Y_{AB} starting from the sampling frames U_v and U_C

$$(7a)\ Y_{AB} = \sum_{k \in U_v} \sum_{j \in U} \frac{1}{L_{vj}} l_{k,j}\, y_j \mathbb{I}\left(L_{Cj} \geq 1\right),$$

$$(7b)\ Y_{AB} = \sum_{k \in U_C} y_k \sum_{j \in U} \frac{1}{L_{Cj}} l_{k,j}\, y_j \mathbb{I}(L_{vj} \geq 1)$$

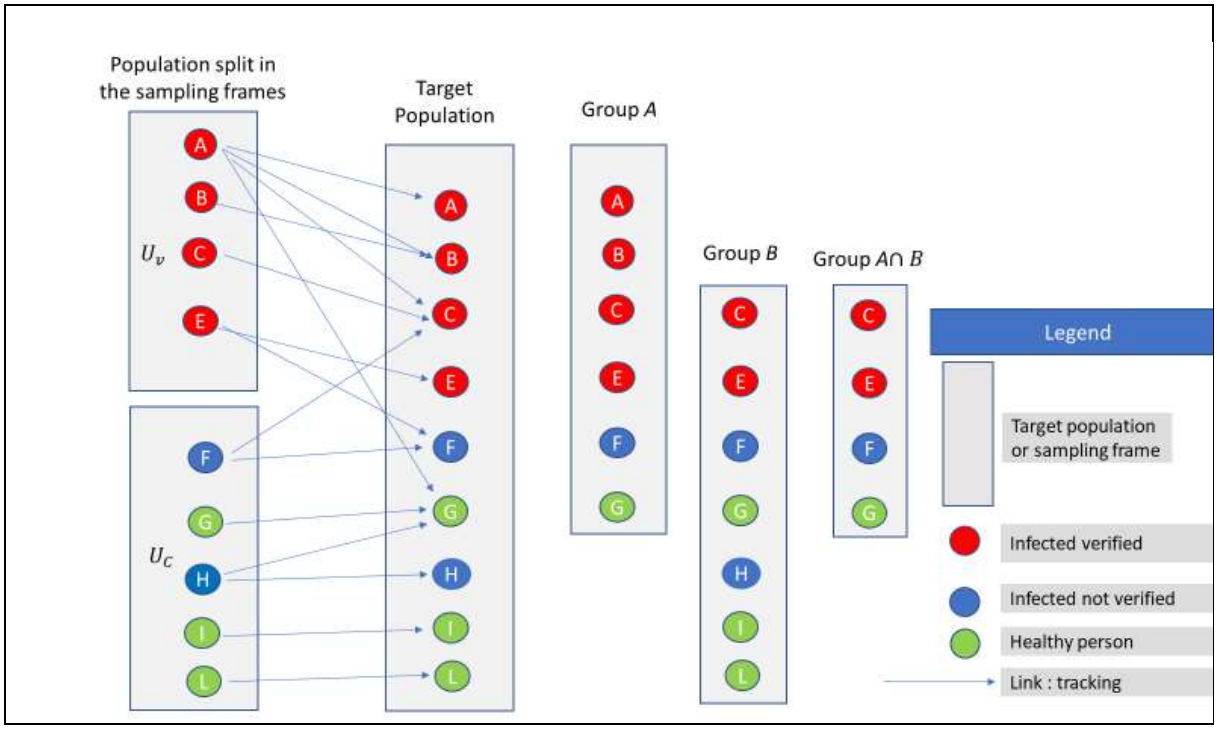
where $\mathbb{I}(A)$ equals 1 if A is true and 0 otherwise. The expressions (7a) and (7.b) are useful in the estimation phase, as illustrated in section 6.3.

Finally, we have

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

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

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



5. The sampling design

5.1. General description of the sampling schemas

Two independent samples, namely S_v and S_C, are selected by the two population subsets, U_v and U_C 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 S_C 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 subset 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 $U_k = \{j \in U: l_{k,j} = 1\}$ of the individual k selected in S_v are tracked going back 14 days.

c) A sample S_{vk} is selected from U_k 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 $S_A = S_v \cup_{k=1}^{\#S_v} S_{vk}$ is formed with a sampling indirect mechanism including people from both S_v (verified infected people) and $\cup_{k=1}^{\#S_v} S_{vk}$ (tracked contacts going back 14 days).

The test to verify the infection is carried out on all the tracked contacts, $\cup_{k=1}^{\#S_v} S_{vk}$. 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 (Scott, 2000 and Yang *et al.,* 2016).

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:

$$\pi_{1i} = m\frac{M_i}{M}, \quad (9)$$

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:

$$\sum_{i\in S_{1v}} \frac{\mathbf{x}_{1v}}{\pi_{1v}} = \sum_{k\in U_{1v}} \mathbf{x}_k, \quad (10)$$

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 (10), 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 (10) 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:

$$\pi_{vk} = m\frac{M_i}{M}\frac{\bar{n}}{M_i} = m\frac{\bar{n}}{M}. \quad (11)$$

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 U_C

In this section, we illustrate the sampling design for the first-time occasion in which we select, independently from S_v, a panel of individuals for estimating the total Y_B. Afterwards, we will monitor these people repeatedly over time.

The operational aspects to be carried out in this first-time occasion are:

a) First of all, a sample S_C is selected without replacement from U_C with inclusion probabilities π_{Ck} $(k = 1,2 \dots, \#U_C)$.

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 U_k are tracked, going 14 days back in time.

c) If $y_k = 1$, a sample S_{Ck} is selected from U_k without replacement and with equal inclusion probability $\pi_{2C|k}$. We adopted for the second stage inclusion probability, $\pi_{2C|k}$, the same notation used for $\pi_{2v|k}$. At end of the whole process, the sample $S_B = S_C \cup_{k=1:y_k=1}^{\#S_C} S_{Ck}$ is formed with an indirect sampling mechanism including people from both S_C (people for which the infection status is not known) and $\cup_{k=1:y_k=1}^{\#S_C} S_{Ck}$ (tracked contacts going back 14 days of the infected people in S_C).

Remark 3. The populations U_v and U_C change as a function of time. The panel can be representative of the shifting population. We discuss this topic later on in section 7. Here, we note that in the subsequent survey occasions, the verified infected people of the panel are automatically captured by the sampling mechanism defined for the population U_v. While, the sample S_C reduces its size, observing only the non-verified infected people. This reduction of the sampling size makes it necessary a regular refresh of the panel over time.

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 U_C could also be identified.

In this informative contexts, one stage sampling design could be carried out with optimal inclusion probabilities π_{Ck} 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:

$$(12) \quad \sum_{k \in S_C} \frac{\mathbf{x}_k}{\pi_{Ck}} = \sum_{k \in S_C} \mathbf{x}_k.$$

Remark 4. 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 5. The two-phase mechanism could be useful if the identification of U_C 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

We first note that in our proposal we sub-sample from the list of contacts. We adopt this choice for controlling the survey costs. However, we could extend the sample on all the set of contacts. Furthermore, if we continue the tracking on the contacts till all the people tracked is not infected, the sample design adopted becomes a classic adaptive schema (Thompson and Seeber, 1996), which can thus be seen as a particular application of our proposal.

Comments and good suggestions in this respect came from a discussion with the Portuguese National Statistical Office (INE) and in particular from Francisco Lima, President, Pedro Campos, Director of the Methodology Department and João Lopes from the same department., with some contributions from Portuguese academia.

Given the complexity of the epidemiology of Covid-19, it may be useful to consider 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 important 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. The population density is also an important factor to control in the phase of sampling design.

6. Sample estimation of the total of infected people

We can compute a direct estimation of the total Y for each time and each territorial unit, as:

(13) $\hat{Y} = \hat{Y}_A + \hat{Y}_B - \hat{Y}_{AB}$,

being

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

where $\hat{Y}_A$ and $\hat{Y}_{AB}^{A}$ are the Generalized Weight Share Method (GWSM, Lavallé, 2007) estimates of the totals Y_A and Y_{AB} derived from the sample S_A; $\hat{Y}_B$ and $\hat{Y}_{AB}^{B}$ are the GWSM

estimates of the totals Y_B and Y_{AB} calculated from the sample S_B and $\hat{Y}_{AB}$ is a convex combination of the GWSM estimates $\hat{Y}_{AB}^{A}$ and $\hat{Y}_{AB}^{B}$, being $0 \leq \alpha \leq 1$.

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

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

Remark 6. 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 interviews.

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

This proof can be found in section 5.1 of Lavallée (2007). Denoting with $E(\cdot)$ the operator of sampling expectation, we have

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

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:

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

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 the expression (20) into equation (19), we finally have:

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

Variance of $\hat{Y}_A$

The main results on this topic can also be found in section 5.1 of Lavallée (2007). On the basis of the theorem on two stage sampling (Cochran, 1977), the variance of $\hat{Y}_A$ can be expressed as follows:

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

In the previous expression the variance is decomposed into the sum of the first stage variance (V_1) and the first stage expectation of the second stage variance (V_2). All the elements of the previous expression can be estimated with standard statistical inferential techniques (see Horvitz and Thompson, 1952 and Kish, 1965).

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

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

$$(22)\ \hat{Y}_B = \sum\nolimits_{k\in S_C}\frac{1}{\pi_{Ck}}y_k\sum\nolimits_{j\in S_{Ck}}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}\, l_{k,j}y_j$$
$$= \sum\nolimits_{k\in S_C}\frac{1}{\pi_{Ck}}\hat{Z}_{Ck}$$

where the term:

$$(23)\quad \hat{Z}_{Ck} = y_k\sum\nolimits_{j\in S_{Ck}}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}\, l_{k,j}y_j,$$

represents the estimate of

(24) $Z_{Ck} = y_k \sum_{j \in U_k} \frac{1}{L_{Cj}}\, l_{k,j} y_j.$

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

To prove the unbiasedness, first of all we have:

(25) $E(\hat{Y}_B) = \sum_{k \in U_C} y_k \sum_{j \in U_k} \frac{E(\delta_{Ck}\, \delta_{2Cj|k})}{\pi_{Ck}\, \pi_{2C|k}} \frac{1}{L_{Cj}}\, l_{k,j} y_j,$

where δ_{Ck} is a dichotomous variable being $\delta_{Ck} = 1$, if $k \in S_C$ and δ_{Ck} =0, otherwise; and $\delta_{2Cj|k}$ is a dichotomous variable being $\delta_{2Cj|k} = 1$, if $y_k = 1 \cap j \in S_{Ck}$ and 0, otherwise.

However, we have:

(26) $E(\delta_{Ck}\, \delta_{2Cj|k}) = E\left[\delta_{Ck} E\left(\delta_{2Cj|k} | \delta_{Ck} = 1\right)\right] = E\left[\delta_{Ck} \pi_{2C|k}\right] = \pi_{Ck}\, \pi_{2C|k}.$

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

$E(\hat{Y}_B) = \sum_{k \in U_C} y_k \sum_{j \in U} \frac{1}{L_{Cj}}\, l_{k,j} y_j\,.$ Q. E. D.

The term L_{Cj} 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. Alternatively, we could determine it by following a *back-tracing process*: if the unit *j* is infected, we should test the infection of Covid-19 on all their contacts.

Variance of $\hat{Y}_B$

The variance may be obtained by simply adapting the expression (21), being:

(27) $V(\hat{Y}_B) = V_1\left(\sum_{k \in S_C} \frac{1}{\pi_{Ck}} Z_{Ck}\right) + \sum_{k \in U_C} \frac{1}{\pi_{Ck}} V_2\left(y_k \sum_{j \in S_{Ck}} \frac{1}{\pi_{2C|k}} \frac{1}{L_{Cj}}\, l_{k,j} y_j\right).$

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

Starting from the expression (7*a*), we obtain the GWSM unbiased estimator of Y_{AB} with the data of the sample S_A, as

(28) $\hat{Y}_{AB}^{A} = \sum_{k \in S_v} \frac{1}{\pi_{vk}} \sum_{j \in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}}\, l_{k,j}\, y_j \mathbb{I}(L_{Cj} \geq 1).$

Starting from the expression (7*b*), we derive the GWSM unbiased estimator of Y_{AB} with the data of the sample S_B, as

(29) $\hat{Y}_{AB}^{B} = \sum_{k \in S_C} \frac{1}{\pi_{Ck}} \sum_{j \in S_C} \frac{1}{\pi_{2C|k}} \frac{1}{L_{Cj}}\, l_{k,j} y_j\, \mathbb{I}(L_{vj} \geq 1).$

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 in the context of multiple frames surveys. Hartley (1962, 1974) proposed choosing α in (14) 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:

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

Unfortunately, the above quantity depends on the variable *y*.

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:

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

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. With respect the expression (28), the GCRE estimator allows estimating the hidden population which would not be visible with either the public health structure nor with the panel survey (e.g. the people died at home) being very difficult to capture with the usual survey techniques. The main problem for adopting the GCRE estimator is that it would require an overlap of the samples of groups A and B.

Remark 8. In Section 8, and in Appendix we see that the maximum of efficiency is gathered from sampling from U_v. At the same time, collecting the value of the variable L_{Cj} could be complex for the need to set-up a following a *back-tracing process*. Thus, a feasible alternative strategy for the estimation of Y could be represented by

$$\hat{Y}_{alt} = \hat{Y}_A + \hat{Y}_C - \hat{Y}_{AC}^A,$$

where $\hat{Y}_c$ is the standard Narain-HT estimate of the total of $\mathcal{y}$ in U_C and $\hat{Y}_{AC}^A$ is the GWSM estimate of the total of $\mathcal{y}$ in the intersection of U_A with U_C obtained by the sample S_A, being

$$\hat{Y}_c = \sum_{k \in S_C} \frac{1}{\pi_{Ck}} y_k,$$

$$\hat{Y}_{AC}^A = \sum_{k \in S_v} \frac{1}{\pi_{vk}} \sum_{j \in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}} l_{k,j}\, y_j \mathbb{I}\left[\left(L_j - L_{Cj}\right) \geq 1\right],$$

in which L_j is the total of contacts of the unit j.

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 $\mathcal{y}$ 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_1 can be estimated, as described in Section 5, using two sources of data, namely:

- the sample $S_{1,v}$, 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 $S_{0,C}$ selected at the time $t = 0$, which is updated over time, since the tests carried out at time $t = 1$ on the individuals of $S_{0,C}$ individuate the *new infected* people of the panel. Then tracking the contacts of the infected people 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.

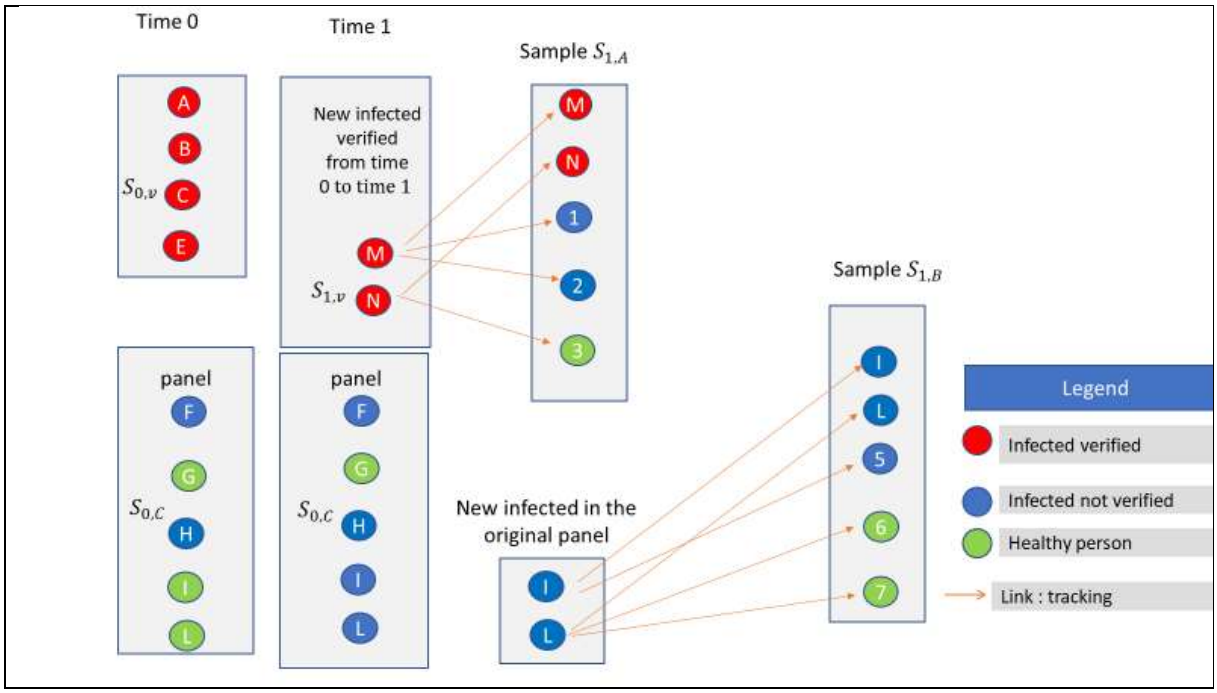
Therefore, we have:

(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 , $\Delta D_{0\to1}$, $\Delta H_{0\to1}$, ΔY_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

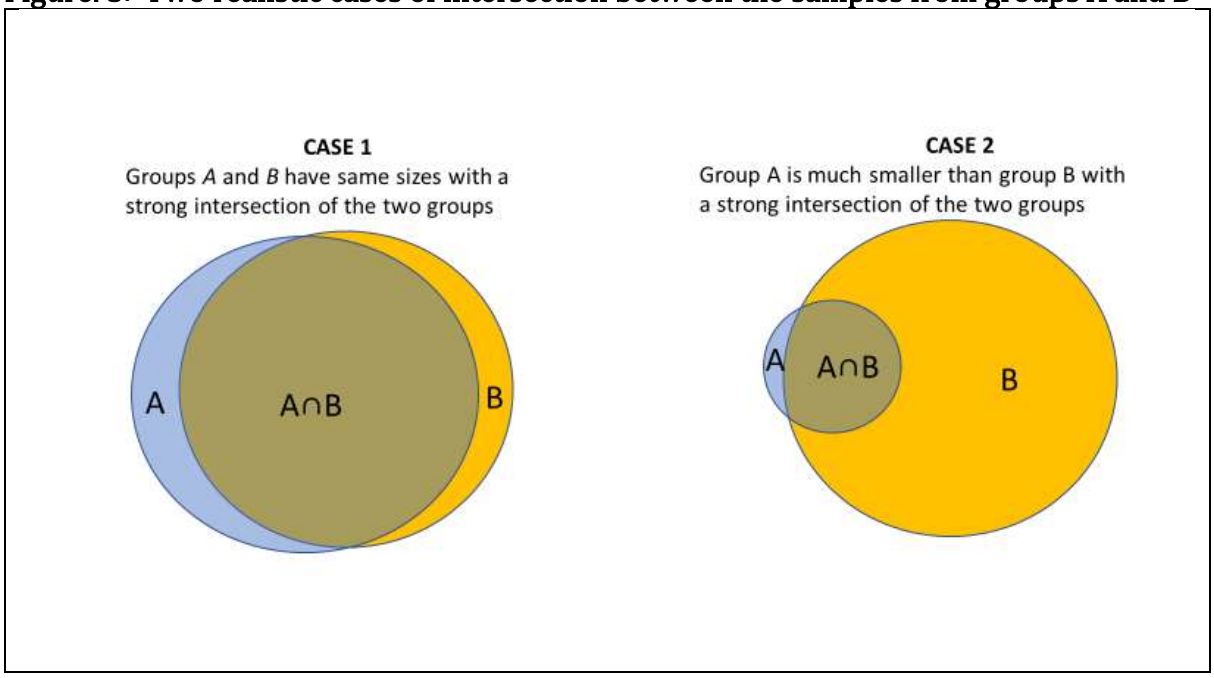


8. A note on the efficiency of the strategy

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

- **Case 1.** Samples from groups A and B have the same size, with a strong intersection between the two groups. This case could characterize the situation in which there is no control of the infection. In this case, the proportions $\gamma_A = Y_{AB}/Y_A$ and $\gamma_B = Y_{AB}/Y_B$ are slightly smaller than 1. Thus, it is possible to consider $\gamma_A \cong 1$ and $\gamma_B \cong 1$.
- **Case 2.** Sample from Group A is much smaller than the sample from group B with a strong intersection between the sample from group A and the intersection between the two samples. In this case, we can consider $\gamma_A \cong 1$ and $\gamma_B \ll 1$. This case could be that in which the infection is controlled by locking down the people.

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



Below, we summarise a result for simple random sampling, which may be useful to understand the efficiency of the strategy better here proposed. We give all the details in the Appendix.

Let us suppose selecting the sample S of size n from U with a SRSWOR design. Let

$$(39)\quad \hat{Y}_{HT,SRS} = \sum_{k=1}^{n} \frac{N}{n} y_k = \sum_{k=1}^{n} \frac{1}{f} y_k$$

be the HT estimate of Y, where $f = n/N$. Let $\mu = Y/N$ be the proportion of infected people in the overall population.

For N large and f small, the Anticipated Variance (AV) of $\hat{Y}_{HT,SRS}$ can be approximated by (Falorsi and Righi, 1915, Appendix 4):

(40) $AV\left(\hat{Y}_{HT,SRS}\right) = \frac{N}{f}\mu(1-\mu).$

Let $U_{yl} = \{k, j: y_k = 1, y_k l_{k,j} = 1; k, j = 1, \dots, N\}$ denote the sub-population of infected people and of those who have had contacts with them. Let $\vartheta = Y/\#U_{yl}$ be the proportion of infected people in U_{yl}, being $\mu \ll \vartheta$. Let us supposed to allocate the sample S proportionally between the two frames U_v and U_C and to select a SRSWOR in each frame. Thus, the sample sizes for U_v and U_C are $P_v n$ and $(1-P_v)n$, respectively.

The GWSM estimates of the totals Y_A , Y_B , Y_{AB} and Y are:

(41) $\hat{Y}_{A,SRS} = \sum_{k=1}^{P_v n} \frac{1}{f} \sum_{j=1}^{N} \frac{1}{L_{vj}}\, l_{k,j}\, y_j \quad , \qquad \hat{Y}_{B,SRS} = \sum_{k=1}^{(1-P_v)n} \frac{1}{f} y_k \sum_{j=1}^{N} \frac{1}{L_{Cj}}\, l_{k,j}\, y_j,$

$$\hat{Y}^{A}_{AB,SRS} = \sum_{k=1}^{P_v n} \frac{1}{f} \sum_{j=1}^{N} \frac{1}{L_{vj}}\, l_{k,j}\, y_j \left(L_{Cj} \geq 1\right),$$

$$\hat{Y}^{B}_{AB,SRS} = \sum_{k=1}^{(1-P_v)n} \frac{1}{f} y_k \sum_{j=1}^{N} \frac{1}{L_{Cj}}\, l_{k,j}\, y_j \left(L_{vj} \geq 1\right).$$

$$\hat{Y}_{SRS} = \hat{Y}_{A,SRS} - \alpha \hat{Y}^{A}_{AB,SRS} + \hat{Y}_{B,SRS} - (1-\alpha)\hat{Y}^{B}_{AB,SRS}.$$

Assuming that the number of contacts, L, is roughly constant in U, the AV of $\hat{Y}_{SRS}$ is

(42) $AV(\hat{Y}_{SRS}) = AV\left(\hat{Y}_{A,SRS} - \alpha \hat{Y}^{A}_{AB,SRS}\right) + AV\left[\hat{Y}_{B,SRS} - (1-\alpha)\hat{Y}^{B}_{AB,SRS}\right]$

where

(43) $AV\left(\hat{Y}_{A,SRS} - \alpha \hat{Y}^{A}_{AB,SRS}\right) \cong \frac{P_v N}{f} \frac{1}{L} \vartheta[(1-\vartheta)(1-2\alpha\gamma_A) + \alpha^2 \gamma_A (1-\gamma_A \vartheta)]$

and

(44) $AV\left(\hat{Y}_{B,SRS} - (1-\alpha)\hat{Y}^{A}_{AB,SRS}\right)$

$$\cong (1-P_v) N \frac{1}{fL\mu} \vartheta[(1-\mu\vartheta) + (1-\alpha)^2 \gamma_B (1-\gamma_B \mu\vartheta) - 2(1-\alpha)\gamma_B (1-\mu\vartheta)].$$

Comparing the expressions (42) and (40), we have that the efficiency of the proposed strategy can be defined as the ratio of the two AVs:

(45) $Eff\left(\hat{Y}_{SRS}\right) = \frac{AV\left(\hat{Y}_{SRS}\right)}{AV\left(\hat{Y}_{HT,SRS}\right)}$

$$= \frac{\frac{1}{L}\vartheta[(1-\vartheta)(1-2\alpha\gamma_A) + \alpha^2\gamma_A(1-\gamma_A\vartheta)]}{\mu(1-\mu)} P_v +$$

(46) $$+ \frac{\frac{1}{L\mu}\vartheta[(1-\mu\vartheta) + (1-\alpha)^2\gamma_B(1-\gamma_B\mu\vartheta) - 2(1-\alpha)\gamma_B(1-\mu\vartheta)]}{\mu(1-\mu)} (1-P_v).$$

Looking at expression (46), we can highlight the following results:

- ✓ The effectiveness of the strategy is maximum for case 1.

- ✓ The efficacy is maximum for the sampling from U_v in which is realistic to have $\frac{1}{L}\vartheta[(1-\vartheta)(1-2\alpha\gamma_A)+\alpha^2\gamma_A(1-\gamma_A\vartheta)] < \mu(1-\mu)$.
- ✓ The efficacy could be lower or null for the sampling from U_C in which the condition

  $$\frac{1}{L\mu}\vartheta[(1-\mu\vartheta)+(1-\alpha)^2\gamma_B(1-\gamma_B\mu\vartheta)-2(1-\alpha)\gamma_B(1-\mu\vartheta)] < \mu(1-\mu),$$

  is not always given.

Thus, a good strategy could be that of oversampling in U_v and having a small sample for U_C.

9. Empirical evaluations of the proposed method: a Monte Carlo study

9.1 Artificial data generation

Since it is not possible, at this stage, to include a numerical illustration using real-life sample data, in this section we report the results of a series of Monte Carlo experiments which justify numerically our proposed ideas and show their statistical performances in an artificial, although as realistic as possible, context.

Before showing our simulation results, we need to clarify the criteria we used in the data generation process and those employed in the generation of the geographical map on which data are observed. This second element is essential, given the peculiar nature of the transmission mechanism which requires physical proximity between infected people.

First of all, in order to simulate an artificial population describing the time evolution of an epidemics, we considered a popular model constituted by a system of six differential equations which, in each moment of time, describe six categories of individuals, namely: the susceptibles (S), those exposed to the virus (E), the infected with symptoms (I), those without symptoms (A) and those that are removed from population either because healed (R) or dead (D). This modelling framework is due to the seminal contribution of Hamer (1906), Kermack and McKendrick (1927) and Soper (1929) and it is often referred to as the "SIR model" from the initials of the categories considered in the first simplified formulation: Susceptibles, Infected and Removed. A comprehensive overview of this model is contained in Cliff et al. (1981). See also Vynnycky and White (2010). Figure 4 describes diagrammatically the transition between the 6 categories. For the data random generation, we assumed that, if infected, a susceptible element of the population (S) will remain in the exposed state (E) for 5 days. After that period the subject can become either infected with symptoms (I) with probability 0.25 or without (asymptomatic; symbol A) with probability 0.75. The asymptomatic will remain infected (and so still able to transmit the virus) for 14 days. After this period all the asymptomatic will be considered healed and will pass to the category removed (R). In contrast, the infected people showing symptoms will be healed with probability 0.85 or die (D) with probability 0.15 (case death rate).

Figure 4. The six basic categories of our simulation model and their transition pattern

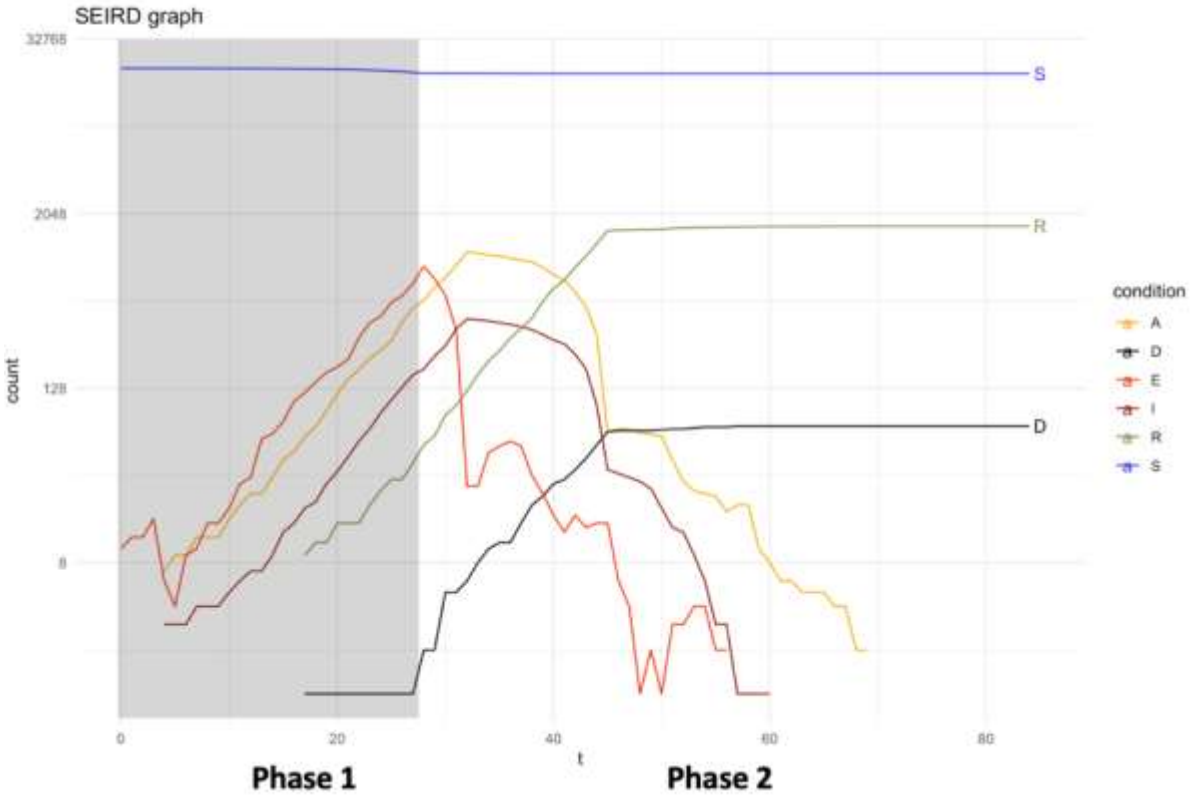
For the map generation we considered a population distributed into 25 spatial units laid on a 5-by-5 regular squared lattice grid. Each square of the grid contains a number of individuals randomly drawn from a uniform distribution ranging between 800 and 1,000. After a simulation exercise with these parameter values, we obtained an artificial population with a total of 22,217 individuals.

This geographical representation is very general in that the map thus generated can represent, e. g., a city divided into blocks or a region divided into smaller spatial union or any other meaningful geographical partition.

The contagion mechanism is favoured by people mobility. In our exercise, we assumed that in any moment of time a certain percentage m of the population can move between the squares. We distinguish two epidemic phases. In Phase 1 people is free to move and such percentage is $m=0.03$, while Phase 2 describes a period of lockdown when mobility is discouraged and $m=0.01$. In particular, we considered Phase 1 involving a period of 4 weeks and Phase 2 related to the period of the 8 subsequent weeks. The commuting during the lockdown period is not only limited by the number of people who move, but also by the extent of their movements. This is a further simulation parameter which is generated by a uniform distribution ranging from -4 to 4 during Phase 1 (thus allowing movements in and out the cells) and between -1 and 1 during the Phase 2.

Given the mobility pattern described above, contagion is determined by the social interaction and the contact opportunities. The number of contacts in each square of the grid is assumed to be determined by a random number drawn from a Poisson distribution with parameter, say c_n, while the number of people involved in the movements is also a Poisson number characterised but a different parameter c_p. Given these assumptions, a contagion occurs in the following way. If in a meeting it is present at least one asymptomatic or an exposed person, i_m susceptibles will be infected moving in the status of the Exposed. In our runs of the simulation, we considered Phase 1 characterised by the following parameters $c_n = 20;\ c_p = 5;\ i_m = 3.$ In contrast, during Phase 2 the three parameters become $c_n = 3;\ c_p = 3;\ i_m = 2$ reflecting the decreased chances of contacts between people. Figure 2 describes the time evolution of the six categories of people in our simulated epidemics. As already said, we consider Phase 1 constituted by 4 weeks (day 1 to day 28) and Phase 2 lasting 8 weeks (day 29 to day 84). Figure 5 shows that. despite the many assumptions that we were forced to include in the simulation, the contagion curves are very similar to those observed worldwide in the recent 2020 SARS-CoV-2 pandemics.

Figure 5. Time evolution of the six categories of people in the simulated epidemics. Phase 1 refers to days 1- 28. Phase 2 refers to days 29-84.



9.2 Simulation results

We present the main results obtained in the simulation exercise. Using the artificial population generated as described in the previous section, we considered the situation of a repeated sampling survey realized in three moments of time, namely at day 15 (during the beginning of Phase 1), at day 25 (still in Phase 1, but in a situation closer to a *plateau*) and at day 35, during the period of lockdown. The situation of the infected in the three moments of time is reported in Table 1 distinguishing between the samples in groups A, B and their intersection (see Figure 3).

Table 1. True simulated population values of the infected (distinguished for the two subpopulations called , Groups A, B and their intersection) at different days

Group considered	Day 15	Day 25	Day 35
Y_A	42	374	1,041
Y_B	126	875	1,432
Y_{AB}	39	372	1,018
Total of infected	**129**	**877**	**1,455**

For group A we fixed the parameter $g = 0.9$ while for group B the parameter f and ν are fixed as follows: $f = 0.; \nu = 12$.

The sample size obtained with such parameters' definition (both excluding and including the contacts) are reported in Table 2 by distinguishing 4 sample situations, namely: (i) A1B2 when both the individuals belonging to group A and their contacts are totally sampled while in group B both the non-infected and all contacts are sampled; (ii) A1B3 when both the individuals belonging to group A and their contacts are sampled while in group B the non-infected are sampled with all contacts with a maximum of $\nu = 12$; (iii) A2B2 when all individuals belonging to group A, but only a subset of their contacts are included in the sample and in group B the non-infected are sampled with all their contacts; and, finally, (iv) A2B3 when all individuals belonging to group A, but only a subset of their contacts are included in the sample while in group B the non-infected are sampled with all their contacts, but only up to a maximum of $\nu = 12$ individuals. Notice that in day 35 we have fewer contacts in the sample than in day 25 due to the lockdown measures considered.

Table 2. Total number of sample units including and excluding the contacts at different days and in the various sampling schemes

Day	Proportion of infected in the population	Sampling scheme	Sampling units without contacts	Sampling units with contacts
15	0.006	A1B2	4,130	4,741
		A1B3	4,130	4,736
		A2B2	4,130	4,741
		A2B3	4,130	4,736
25	0.042	A1B2	4,198	7,650
		A1B3	4,198	7,634
		A2B2	4,198	7,650
		A2B3	4,198	7,634
35	0.070	A1B2	4,361	7,545
		A1B3	4,361	7,514
		A2B2	4,361	7,545
		A2B3	4,361	7,514

The main results of the simulation are reported in Table 3 which shows that in all sampling settings the relative bias is very small and our estimators outperform dramatically the simple random sampling in terms of efficiency (the ratio of the standard error of the proposed estimator, computed by the simulation, over that of the HT estimator of a simple random sampling without replacement). In particular, the relative bias is of the order of 0.01 % during Phase 1, while during Phase 2 it depends on the sampling scheme adopted with a greater precision when both the individuals belonging to group A and their contacts are included in the sampling. In contrast, the relative bias obviously increases when only a subset of the contacts is observed. Furthermore, our method outperforms the simple random sample also in terms of efficiency. Similar to the case of the bias, the relative advantage with respect to simple random sample is greater in the case of the A1 sample scheme when all selected individuals and their contacts are included in the sample while it is lower in the case of A2 when only a subset of them is

observed. Moreover, Table 3 also displays a decrease in the relative advantage of our method in the day 35 wave where, due to the lockdown restrictions, the number of contacts is much more limited.

The results presented here depend essentially on the particular setting of the (many) parameters involved in the simulation which describe different epidemic evolutions. To mitigate such a subjectivity, we also run many other Monte Carlo experiments using different parameter values. Although available upon request to the authors, these results are not reported here for the sake of succinctness. However, they all confirm the same features of a very low relative absolute bias involved with our proposed method and of its superiority with respect to the simple random scheme in terms of efficiency.

Table 3: Main results of the simulation study for the various sampling schemes at different days

Days (1)	Percentage of infected people in the population (2)	Sampling scheme (3)	Estimated total infected (average on 500 simulations) (4)	α^* (5)	Standard error (6)	Coefficient of variation $\frac{(6)}{(4)}x100$ (7)	True population value (from Table 1) (8)	Relative absolute bias $\frac{\lvert(8)-(4)\rvert}{(4)}$ (9)	Relative efficiency with respect to the simple random sample without contacts (10)	Relative efficiency with respect to the simple random sample with contacts (11) [a]
15	0.0058	A1B2	128.99	0	0.01	0.09	129	0.0001	0.0006	0.0006
		A1B3	128.99	0	0.01	0.09	129	0.0001	0.0006	0.0006
		A2B2	128.75	0	1.56	0.97	129	0.0019	0.0659	0.0718
		A2B3	128.75	0	1.56	0.97	129	0.0019	0.0659	0.0718
25	0.0394	A1B2	876.83	0	0.17	0.05	877	0.0001	0.0027	0.0041
		A1B3	876.84	0	0.16	0.05	877	0.0001	0.0027	0.0040
		A2B2	876.90	1	0.48	0.08	877	0.0001	0.0080	0.0120
		A2B3	877.02	0.48	2.43	0.18	877	0.0000	0.0403	0.0605
35	0.0654	A1B2	1,455	0	0	0	1.455	0	0	0
		A1B3	1,455	0	0	0	1.455	0	0	0
		A2B2	1.461.59	0	9.78	0.21	1.455	0.0045	0.1310	0.1895
		A2B3	1.461.59	0	9.78	0.21	1.455	0.0045	0.1310	0.1895

(a)The relative efficiency is computed as the ratio of the standard error of the proposed estimator (computed by the simulation) over that of the HT estimator of a simple random sampling without replacement.

10. 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 pandemics 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 epidemics 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:

- 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, also taking into account different target variables in the different stages of development of the epidemic. It contributes to implement a statistical surveillance system on the epidemic to be integrated with the existing systems managed by the health authorities;
- Statistical quality. In the paper the properties of the estimators have been formally proved and confirmed analysing the results of a set of Monte Carlo experiments; they guarantee their reliability in terms of unbiasedness and higher efficiency with respect to the simple random sample;
- 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 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; Lavallée and Rivest, 2012) which would require an overlap of the samples of groups *A* and *B*. A further improvement to be explored could derive from applying the Dorfmann procedure (Dorfmann, 1943) to reduce number of the tests and cost.

In addition to the methodological advances, 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. To this aim, it could be interesting to study the possibility of considering, within our framework, the proposal developed by Saunders-Hastings *et al.* (2017) who address the problem *via* a model approach. The need to monitor over time the pandemics should represent the drive for building an integrated surveillance system. This should merge within a unified database three different pieces of

information: (i) the information collected by the administrative institutions when receiving and treating individuals that have turned to the healthcare system; together, (ii) the statistical information collected on purpose with the aim to accurately measure the diffusion of the infection and, finally, (iii) the data obtained through new sources for tracking the movements of people and of their contacts.

A third line of extension of our proposal concerns the operational point of view. Indeed the sample design described in detail in Section 5 should be accompanied by the rapid definition of some key points:

- a *control room* that ensures the necessary inter-institutional collaboration to guide field operations (Health Authorities, at national and regional level, Statistical Offices, others);
- an effective information campaign among the population to promote their participation; the legal framework to assure the collection and the analysis of personal data;
- the medical testing procedure to consider for the selected population (swabs, blood testing and DNA mapping);
- 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; and we proved its feasibility defining the previous key point to estimate times and costs (Alleva et al., 2020)[12]. 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 Covid-19 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.

[12] The sample size to assure a certain level of accuracy of the estimates depends on the Base Rate of infection. The unit cost of the swab and serological administering relies on the level of involvement in the survey of the public health authorities. The total cost depends on the length and the periodicity of the panel survey. For Italy we estimated the cost of data collection at national and regional level (21 regions), in case of 3 months of monitoring, panel survey every 15 days and a Base Rate of infection 0,04. Concerning Groups A and B, the sample sizes are 1,000 and 1,200 units, that implies 6,000 and 7,200 swabs and a total cost of 210,000 and 252,000 euros.

Acknowledgements. We are very grateful to Mike Hidiroglou, Pierre Lavallée and Giovanna Ranalli for the challenging discussion, careful reading and the useful suggestions which have helped us to improve the quality of our proposal.

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., G. Arbia, P. D. Falorsi, G. Pellegrini, A. Zuliani. 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.

Alleva G. 2020. Contributo per la 12° Commissione permanente Igiene e sanità del Senato della Repubblica, Roma, 27 may 2020. 2020.https://www.senato.it/application/xmanager/projects/leg18/attachments/documento_evento_procedura_commissione/files/000/135/501/GIORGIO_ALLEVA.pdf

Alleva G., A. Zuliani. 2020, Coronavirus: chiarezza sui dati, *Bancaria,* ISSN: 0005-4623.

Aguilar, J. B., J. S. Faust, L. M. Westafer and J. B. Gutierrez. 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 G. Lafratta. 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 G. Lafratta. 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. *Journal of Ornitholo*gy, 152**,** 435–444. https://doi.org/10.1007/s10336-010-0583-z

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

Cliff, A. D., Haggett, P., Ord, J. K. and Verfey, F. R. (1981) *Spatial Diffusion: an Historical Geography of Epidemics in an Island Community*. Cambridge University Press.

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

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

Deville, J.-C. and Y. Tillé. 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.

Di Gennaro, Splendore and Luca. 2020. Random testing, quality of data and lack of information: COVID-19. Available at: https://medium.com/data-policy/random-testing-quality-of-data-andlack-of-information-covid-19-a6e09a398d1d

Dorfman, R. 1943. The Detection of Defective Members of Large Populations, *The Annals of Mathematical Statistics*, 14 (4): 436–440, doi:10.1214/aoms/1177731363

Falorsi P. D., P. Righi. 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., P. Righi, P. Lavallée. 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, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, C. Fraser. 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, *Science* 31 Mar 2020, DOI:10.1126/science.abb6936.
https, //science.sciencemag.org/content/early/2020/03/30/science.abb6936 .

Fuggetta M. 2020. Testing for the Base Rate.
massimofuggetta.com/2020/04/28/testing-for-the-base-rate/.

Giuliani, D., M. M. Dickson, G., Espa, and F. Santi. 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, Z. Zheng, S. Liu, X. Chen, A. Sela, J. Li, D. Li, S. 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 .

Hamer W. H. (1906), *Epidemic diseases in England*, Lancet, 1

Hansen N.H., N.W. Hurwitz, W.G. Meadow. 1953. *Sample Survey Method and Theory*. Wiley, New York.

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 D.L. Thompson. 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/.

Ioannidis, J. 2020, A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. Available at:

https://www.statnews.com/2020/03/17/afiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-withoutreliable-data/

Istituto nazionale di statistica (Istat). Primi risultati dell'indagine di sieroprevalenza sul SARS-CoV-2 https://www.istat.it/it/files//2020/08/ReportPrimiRisultatiIndagineSiero.pdf

Kermack, W.O. and McKendrick, A. G. (1927) A contributions to the mathematical theory of epidemics" Proceedings pf the Royal society London, series A, 115, 700-721.

Kiesl, H.. 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., L. P. Rivest. 2012. Capture–Recapture Sampling and Indirect Sampling. *Journal of Official Statistics*, Vol. 28, No. 1, 2012, pp. 1–27.

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

Leung, G. and K. Leung. 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., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. 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., K. Kagaya, A., Zarebski and G. Chowell. 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., K. Katsushi, A. Zarebski, 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 V. K. Sethi. 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.

Office for National Statistics (ONS). Coronavirus (COVID-19) Infection Survey pilot: England and Wales, 14 August 2020.
https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales14august2020

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., A. Keshet, S. Shilo, A. Gavrieli, T. Bauman, O. Cohen, R. Balicer, B. Geiger, Y. Dor, E. 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.

Saunders-Hastings, P., B. Q. Quinn Hayes, R. Smith, D. Krewski. 2017. *Control strategies to protect hospital resources during an influenza pandemic*.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179315

Scott J., 2000. *Social Network Analysis. A Handbook*, London, Sage Publications.

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

Soper H. E. (1929), Interpretation of periodicity in disease prevalence, *Journal of the Royal Statistical Society*, *A*, 92, 34-73

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

Sun., K., J. Chen and C. 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 .

Tillé Y. and M. Wilhelm. 2017. Probability Sampling Designs: Principles for Choice of Design and Balancing. *Statistical Science*. Volume 32, Number 2 (2017), 176-189.

Thompson, S. K., G. A. F. Steven. 1996. Adaptive sampling, Wiley Blackwell, N-Y.

Vynnycky, E.; White, R. G., eds. (2010). *An Introduction to Infectious Disease Modelling*. Oxford: Oxford University Press.

Yang S., F.B. Keller, L. Zheng.2016. *Social Network Analysis: Methods and Examples*, Sage Publications, London.

Yelin, I. , Noga Aharony, Einat Shaer-Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Kuzli, A., N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, R. 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.

APPENDIX

For the estimator (39), the following model can be assumed,

$$(A1)\quad E_M(y_k)=\mu,\qquad V_M(y_k)=\mu(1-\mu)\ \text{ for } k=1,\dots,N,$$

where E_M and V_M denote the model expectation and variance. According to the above the Anticipated Variance of the estimate $\hat{Y}_{HT,SRS}$ may be defined as (Falorsi and Righi, 1915, Appendix 4):

$$(A2)\quad AV\left(\hat{Y}_{HT,SRS}\right)=\frac{N}{N-1}\sum_{k=1}^{N}\left(\frac{N-n}{n}\right)V_M(y_k)=\frac{N}{N-1}\sum_{k=1}^{N}\frac{N}{n}\left(\frac{N-n}{N}\right)\mu(1-\mu)$$

$$(A3)\qquad \cong\sum_{k=1}^{N}\frac{N}{n}\mu(1-\mu)\ =\sum_{k=1}^{N}\frac{1}{f}\mu(1-\mu)=\frac{N}{f}\mu(1-\mu).$$

The approximation $(A3)$ holds for N large and f small.

For the estimators (41), we can introduce the following model

$$(A4)\quad E_M(y_k)=\vartheta,\qquad V_M(y_k)=\vartheta(1-\vartheta)\ \text{ for } k\in U_{yl}.$$

Furthermore, for the variables $y_j\ (L_{Cj}\geq 1)$ and $y_j\ (L_{vj}\geq 1)$ we can adopt the hypothesis that the probabilities

$$(A5)\quad P\left(L_{Cj}\geq 1|k\in U_v\cap l_{k,j}=1\right)\cong\gamma_A$$

$$P\left(L_{vj}\geq 1|k\in U_C\cap y_{k,j}\,l_{k,j}=1\right)\cong\gamma_B$$

are roughly constant. Then we may derive the following models

$$(A6)\quad E_M\left[y_j\ (L_{Cj}\geq 1)\right]=\vartheta\gamma_A,\ V_M\left[y_j\ (L_{Cj}\geq 1)\right]=\vartheta\gamma_A(1-\vartheta\gamma_A)\ \text{ for}k\in U_v\cap l_{k,j}=1,$$

$$E_M[y_j\ (L_{vj} \geq 1)] = \vartheta\gamma_B, V_M[y_j\ (L_{vj} \geq 1)] = \vartheta\gamma_B(1-\vartheta\gamma_B) \text{ for } k \in U_C \cap y_{k,j}\ l_{k,j} = 1.$$

Let us assume that number of contacts, L, are roughy constant in U. Then, the total number of contacts deriving from the two the frames (U_v and U_C) are

$$TL_v = \sum_{k=1}^{P_vN}\sum_{j=1}^{N} l_{k,j} = P_vNL\ , \qquad TL_C = \sum_{k=1}^{(1-P_v)N} y_k \sum_{j=1}^{N} l_{k,j} \cong \mu(1-P_v)NL.$$

By considering the reasonable assumptions of an uniform distribution of the contacts among the units, we have:

$$L_{vj} \cong \frac{P_vNL}{P_vN} = L \text{ and } L_{Cj} \cong \frac{\mu(1-P_v)NL}{(1-P_v)NL} = L\mu.$$

Adopting the model $(A1)$ in the case we do not know if the unit $k \in U_{yl}$ and the models $(A4)$ for $k, j \in U_{yl}$, then the estimates $\hat{Y}_{A,SRS}$ and $\hat{Y}_{B,SRS}$ can be approximated by

$$(A7)\quad \hat{Y}_{A,SRS} \cong \sum_{k=1}^{P_vn}\frac{1}{f}\sum_{j=1:j\in U_k}^{L}\frac{1}{L}\ y_j\ , \qquad \hat{Y}_{B,SRS} = \sum_{k=1}^{(1-P_v)n}\frac{1}{f}y_k\sum_{j=1:j\in U_k}^{L}\frac{1}{L\mu}\ l_{k,j}\ y_j,$$

$$\hat{Y}^{A}_{AB,SRS} = \sum_{k=1}^{P_vn}\frac{1}{f}\sum_{j=1:j\in U_k}^{L}\frac{1}{L}y_j\ \gamma_A\ ,$$

$$\hat{Y}^{B}_{AB,SRS} = \sum_{k=1}^{(1-P_v)n}\frac{1}{f}y_k\sum_{j=1:j\in U_k}^{L}\frac{1}{L\mu}y_j\ \gamma_B.$$

Then, for the estimates derived from S_A, we have:

$$(A8)\quad AV(\hat{Y}_{A,SRS}) \cong \sum_{k=1}^{P_vN}\frac{1}{f}\sum_{j=1:j\in U_k}^{L}\frac{1}{(L\vartheta)^2}\ V_M\ (y_j) = \sum_{k=1}^{P_vN}\frac{1}{f}\sum_{j=1:j\in U_k}^{L}\frac{1}{L^2}\ \vartheta(1-\vartheta)$$

$$= \sum_{k=1}^{P_vN}\frac{1}{f}\frac{L}{(L\vartheta)^2}\vartheta(1-\vartheta) = \frac{P_vN}{f}\frac{1}{L}\vartheta(1-\vartheta),$$

$$AV(\hat{Y}^{A}_{AB,SRS}) = \sum_{k=1}^{P_vN}\frac{1}{f}\sum_{j=1:j\in U_k}^{L}\frac{1}{L^2}\ \gamma_A\vartheta(1-\gamma_A\vartheta) = \frac{P_vN}{f}\frac{1}{L}\gamma_A\vartheta(1-\gamma_A\vartheta)$$

$$ACov(\hat{Y}_{A,SRS}, \hat{Y}^{A}_{AB,SRS}) = \sum_{k=1}^{P_vN}\frac{1}{f}\frac{L}{L^2}Cov_M[y_j\ (L_{Cj} \geq 1), y_j] = \frac{P_vN}{f}\frac{1}{L}\gamma_A\vartheta(1-\vartheta).$$

Putting together the above results we have:

$$(A9)\quad AV(\hat{Y}_{A,SRS} - \alpha\hat{Y}^{A}_{AB,SRS}) = AV(\hat{Y}_{A,SRS}) + \alpha^2 AV(\hat{Y}^{A}_{AB,SRS}) - 2\alpha ACov(\hat{Y}_{A,SRS}, \alpha\hat{Y}^{A}_{AB,SRS})$$

$$\cong \frac{P_vN}{f}\frac{1}{L}\vartheta(1-\vartheta) + \alpha^2\frac{P_vN}{f}\frac{1}{L}\gamma_A\vartheta(1-\gamma_A\vartheta)$$

$$- 2\alpha\frac{P_vN}{f}\frac{1}{L}\gamma_A\vartheta(1-\vartheta)$$

$$= \frac{P_vN}{f}\frac{1}{L}\vartheta[(1-\vartheta)(1-2\alpha\gamma_A) + \alpha^2\gamma_A(1-\gamma_A\vartheta)].$$

In order to evaluate the anticipated variances and covariances for the estimates derived from S_B, we have to preliminarly consider this results:

$(A10)$ $E_M(y_k y_j) = \Pr(y_k = 1)\, E_M(y_j | y_k = 1) + [1 - \Pr(y_k = 0)]0 = \mu\vartheta$ for $k \in U_C \cap j \in U_k$,

$$V_M(y_k y_j) = E_M\left[(y_k y_j)^2\right] - \left[E_M(y_k y_j)\right]^2 = \mu\vartheta(1 - \mu\vartheta). \quad \text{for } k \in U_C \cap j \in U_k,$$

$$V_M\left[y_k y_j \left(L_{vj} \geq 1\right)\right] = \mu\vartheta\gamma_B(1 - \gamma_B\mu\vartheta),$$

$$Cov_M\left[y_k y_j \left(L_{vj} \geq 1\right), y_k y_j\right] = \mu\vartheta\gamma_B(1 - \mu\vartheta).$$

Then, we have

$$(A11)\ AV\left(\hat{Y}_{B,SRS}\right) \cong \sum_{k=1}^{(1-P_v)N} \frac{1}{f} \sum_{j=1:j\in U_k}^{L} \frac{1}{(L\mu)^2}\, V_M(y_k y_j) = (1 - P_v)N\frac{1}{fL\mu}\vartheta(1 - \mu\vartheta),$$

$$AV\left(\hat{Y}^B_{AB,SRS}\right) \cong (1 - P_v)N\frac{1}{fL\mu}\vartheta\gamma_B(1 - \gamma_B\mu\vartheta)$$

$$ACov\left(\hat{Y}_{B,SRS}, \hat{Y}^B_{AB,SRS}\right) \cong (1 - P_v)N\frac{1}{fL\mu}\vartheta\gamma_B(1 - \mu\vartheta).$$

Putting together the above results we have:

$$(A12)\quad AV\left(\hat{Y}_{B,SRS} - (1 - \alpha)\hat{Y}^A_{AB,SRS}\right)$$
$$= AV\left(\hat{Y}_{B,SRS}\right) + (1 - \alpha)^2 AV\left(\hat{Y}^B_{AB,SRS}\right) - 2(1 - \alpha)ACov\left(\hat{Y}_{B,SRS}, \alpha\hat{Y}^B_{AB,SRS}\right)$$
$$= (1 - P_v)N\frac{1}{fL\mu}\vartheta[(1 - \mu\vartheta) + (1 - \alpha)^2\gamma_B(1 - \gamma_B\mu\vartheta) - 2(1 - \alpha)\gamma_B(1 - \mu\vartheta)].$$