Adaptive cluster sampling-based design for estimating COVID-19 cases with random samples

Girish Chandra[1,]*, Neeraj Tiwari[2] and Raman Nautiyal[2]

[1]Division of Forestry Statistics, Indian Council of Forestry Research and Education, Dehradun 248 006, India
[2]Department of Statistics, Kumaun University, SSJ Campus, Almora 263 601, India

During the COVID-19 pandemic, testing of all persons except those who are symptomatic, is not feasible due to shortage of facilities and staff. This article focuses on estimating the number of COVID-19-positive persons over a geographical domain. The Horvitz–Thompson and Hansen–Hurwitz type estimators under adaptive cluster sampling-based design have been suggested. Two case studies are discussed to demonstrate the performance of the estimators under certain assumptions. Advantages and limitations are also mentioned.

Keywords: Adaptive cluster sampling, COVID-19, pandemic, precise estimation, random samples.

ADAPTIVE cluster sampling (ACS), introduced by Thompson[1], is an efficient, two-phase sampling procedure for estimating totals and means of rare and clustered populations. It is based on a reasoning that rare species are found in clusters or hotspots over a given area. In the first phase, an initial sample is selected by some ordinary sampling scheme, like simple or stratified random sampling and the units which satisfy a previously specified condition *C* (say, presence of at least one rare species) are identified. In the second phase, units in the neighbourhood of identified units are also added to the sample. If any of the newly added units further satisfy *C*, units in their neighbourhoods are also added until the sample includes all the neighbours of any unit satisfying *C*. Thus, appropriate selection of *C* plays an important role. Inappropriate selection of *C* may result in under- or over-sampling, leading to the possibility of imprecise estimation of the population parameters. In order to overcome this problem, Thompson[2] and Chandra *et al.*[3] suggested the use of sample order statistics. In some situations there is no choice for *C*.

Coronaviruses belong to a large family of viruses which causes respiratory infections. The COVID-19 virus has caused deaths in about 3% to 20% of infected persons. The situation is so acute that the World Health Organisation (WHO) declared it as a pandemic on 11 March 2020. As on 11 August 2020, about 215 countries were affected by this virus. In India, almost all the states have been affected, with Maharashtra leading the tally. It is estimated that the incubation period of the virus ranges from 2 to 14 days. All governments are taking steps like complete lockdown for a certain time-period, curfew in many areas, sealing of hotspot areas, etc.

The estimation of COVID-19 cases in an area is the greatest challenge. Besides the reported cases by the authorities, there are many people who either do not want to come forward for tests, or they do not have any symptoms even though they are infected. In such situations, responsible organizations like ICMR, are advising the Governments, both central and state, to select people for testing using simple random sampling without replacement. Many other methods are being used to estimate the actual number of COVID-19 cases in the country; however, the precise estimation is still a challenge. The method based on ACS has been suggested in the present study to provide precise estimates of the number of infected people.

ACS-based design for assessing COVID-19

Suppose there is a population of N_t persons (labelled 1, 2, …, N_t) over a domain (say, a state) at time t ($t = 0, 1, 2, ...$) which is suspected to be infected. If any sub-domain has atleast one COVID-19-positive person, the whole sub-domain is considered as a part of the population under interest. Here, the main variable of interest is of dichotomous type – affected or not, and is defined as

$$y_{it} = 1, \text{ if the } i\text{th person is COVID-19-positive at time } t,$$
$$y_{it} = 0, \text{ otherwise.}$$

We need to estimate the total of y-values at different points of time $\hat{Y}_t = N_t \hat{\mu}_t$, where

$$\hat{\mu}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} y_{it},$$

the population mean at time t.

*For correspondence. (e-mail: gchandra23@yahoo.com)

Since the virus that causes COVID-19 is mainly transmitted through contact with respiratory droplets and now WHO has confirmed its transmission through air although not as extensively, explaining the exact definition of a neighbourhood is difficult. The 'neighbourhood' of two persons does not depend on their physical distance since they can move from a particular place after contacting and spreading the virus. In the present case, the neighbourhood of each person follows a symmetric relation. That is, if person i is in the neighbourhood of a person j, then person j is also in the neighbourhood of person i. A neighbourhood might be the person itself together with all those who come in contact with him/her during few days before (in this case we take the quarantine period of 14 days) at least once. The size of the neighbourhood of each individual may vary from person to person. Thus, the neighbourhood does not depend on the y-values of population, i.e. whether the person is positive or not.

The first phase of this ACS based design starts with drawing a random sample of n persons from the N_t persons, testing all of them for possible occurrence of COVID-19 virus, and identifying the positive persons and isolating them, while simultaneously marking those negative individuals who have been exposed to the COVID-19 virus and therefore have a potential to carry it and, in later stages, develop symptoms associated with the illness, have recent travel history, etc. and quarantine them for 14 days and testing thereafter for possible corona positive.

In the second phase, the neighbourhood of all COVID-19-positive persons is added in the sample and tested. If any person is found to be positive, then his/her entire neighbourhoods are tested, and the remaining persons are quarantined for 14 days. This process is continued until a cluster is obtained that has a boundary of persons found negative after testing. If any of the quarantined persons are found positive, then this adaptive scheme is applied to them until no person is found positive for COVID-19. These boundary persons of each cluster are called edge persons. The final sample then consists of n (not necessarily distinct) clusters, one for each person selected in the initial sample of size n.

There are two types of samples, one is the initial sample s_1 of size n drawn in the first phase and the other is the final sample $s_t = \{s_1, s_{2(t)}\}$ which also includes the initial sample s_1, where $s_{2(t)}$ is the set of distinct unordered labels from the remainder of the sample s_t. N_t and s_t are random variables varying over t.

For finding the estimators for any characteristic, we use the following process. N_t persons are partitioned into K_t disjoint subsets, termed as networks. The network $A_i(t)$ for person i of s_1 is defined to be the cluster generated by i at time t, but with the edge persons removed. If any person from s_1 belongs to a particular network, then all the persons of that network are to be included in s. The definition of cluster is

Cluster = Network + Edge persons.

If person i is the only COVID-19-positive person at time t in a cluster, then $A_{i(t)}$ consists of just person i and forms a network of size one. Any negative person is treated as a network of size one, as this selection does not lead to the inclusion of any other person. Thus, all clusters of size one are also networks of size one. It should be noted that any cluster consisting of more than one person can be split into a network and further networks of size one, as each edge person is a network of size one.

Let m_{it} denote the number of persons in the network $A_{i(t)}$. If $m_i(t) > 1$, we confirm that at least one person is COVID-19-positive in $A_{i(t)}$. If $m_{i(t)} = 1$, then either person i is part of the initial sample but negative, or part of s_2 at time t. Let $a_{i(t)}$ denote the total number of persons in the networks in which the ith person is an edge person at t. If person i is positive, then $a_i(t) = 0$. The probability that the person i is included in s_1 at time t is given by

$$p_{i(t)} = \frac{m_{i(t)} + a_{i(t)}}{N_t}.$$

The probability that the ith person is included in s at time t is

$$\pi_{i(t)} = 1 - \left[\frac{\binom{N_t - m_{i(t)} - a_{i(t)}}{n}}{\binom{N_t}{n}}\right] = 1 - \left[\frac{\binom{N_t - N_t p_{i(t)}}{n}}{\binom{N_t}{n}}\right]. \tag{1}$$

Although $m_{i(t)}$ is known for all the values of i in s, some of the $a_{i(t)}$s may be unknown. These are cases when person i is an edge person for some clusters in the sample. In such cases, all the clusters to which the person belongs would not generally be sampled; so the $a_{i(t)}$ are unknown for those clusters. These issues were solved by Thompson[1] by dropping $a_{i(t)}$ from $\pi_{i(t)}$ and considering the partial inclusion probability. The partial inclusion probability can be interpreted as the probability that the initial sample s_1 intersects $A_{i(t)}$. This intersection probability is denoted by $\alpha_{k(t)}$, for each person i in the kth network at time t. Hence

$$\alpha_{k(t)} = 1 - \left[\frac{\binom{N_t - m_{i(t)}}{n}}{\binom{N_t}{n}}\right]. \tag{2}$$

Equation (2) shows that for large values of $m_{i(t)}$, i.e. the network having more COVID-19 cases

$$\binom{N_t - m_{i(t)}}{n}$$

is small and hence intersection probability $\alpha_{k(t)}$ is more. It is expected that with the increase of $m_{i(t)}$, N_t increases.

Estimators for population total

In order to estimate the population total, Horvitz–Thompson (HT)-type and Hansen–Hurwitz (HH)-type estimators are considered. They are the same type of unbiased estimators proposed by Thompson[1] for population mean which do not depend upon any assumptions regarding population parameters. However, certain restrictions are imposed in the present study. For example, the population size refers to the meaning given earlier and follows the strict inequality $C = \{y : y \geq 1\}$. Besides, the proposed design has certain deviations from the conventional ACS in the sense that the networks and their sizes, population size and intersection probabilities vary over time t; however, the initial sample size n is fixed.

HT-type estimator

In this type of estimator[1], all those negative persons are ignored who were not included in s_1. Thus, the samples of n networks (not necessarily distinct) rather than the n clusters are taken for estimating population total. The HT-type estimator of population total at time t, $\hat{Y}_{\text{HT}(t)}$ can be written in terms of distinct networks generated by s_1 as

$$\hat{Y}_{\text{HT}(t)} = \sum_{k=1}^{K_t} \frac{y^*_{k(t)} J_{k(t)}}{\alpha_{k(t)}},$$

where $y^*_{k(t)}$ is the total positive persons in the kth network, K is the total number of distinct networks in the population and $J_{k(t)}$ takes a value of 1 (with probability $\alpha(k_t)$) if s_1 intersects the network k at t, and 0 otherwise. If K'_t denotes the number of distinct networks in the sample s_t at t, then

$$\hat{Y}_{\text{HT}(t)} = \sum_{k=1}^{K'_t} \frac{y^*_{k(t)}}{\alpha_{k(t)}}. \tag{3}$$

It is to be noted that the estimator (eq. (3)) is unbiased and its variance is written as

$$\text{Var}(\hat{Y}_{\text{HT}(t)}) = \left[\sum_{j=1}^{K_t}\sum_{k=1}^{K_t} y^*_{j(t)} y^*_{k(t)} \left(\frac{\alpha_{jk(t)} - \alpha_{j(t)}\alpha_{k(t)}}{\alpha_{j(t)}\alpha_{k(t)}}\right)\right], \tag{4}$$

where

$$\alpha_{jk(t)} = 1 - \frac{\left[\binom{N_t - m_{j(t)}}{n} + \binom{N_t - m_{k(t)}}{n} - \binom{N_t - m_{j(t)} - m_{k(t)}}{n}\right]}{\binom{N_t}{n}},$$

with $\alpha_{jj(t)} = \alpha_{j(t)}$.

Now the unbiased estimator of $\text{Var}(\hat{Y}_{\text{HT}(t)})$ is

$$\hat{\text{Var}}(\hat{Y}_{\text{HT}(t)}) = \left[\sum_{j=1}^{K_t}\sum_{k=1}^{K_t} y^*_{j(t)} y^*_{k(t)} \left(\frac{\alpha_{jk(t)} - \alpha_{j(t)}\alpha_{k(t)}}{\alpha_{jk(t)}\alpha_{j(t)}\alpha_{k(t)}}\right) J_{j(t)} J_{k(t)}\right].$$

In the present case, generating all distinct networks for the population is difficult than forming distinct networks in the sample. Therefore, an unbiased estimator of the variance of $\hat{Y}_{\text{HT}(t)}$ is[1]

$$\hat{\text{Var}}(\hat{Y}_{\text{HT}(t)}) = \left[\sum_{j=1}^{K'_t}\sum_{k=1}^{K'_t} y^*_{j(t)} y^*_{k(t)} \left(\frac{\alpha_{jk(t)} - \alpha_{j(t)}\alpha_{k(t)}}{\alpha_{jk(t)}\alpha_{j(t)}\alpha_{k(t)}}\right)\right], \tag{5}$$

provided that none of the joint probabilities $\alpha_{jk(t)}$ is zero.

The estimator $\hat{Y}_{\text{HT}(t)}$ should have low variance when the network totals $y^*_{k(t)}$s are proportional to the intersection probability $\alpha_{k(t)}$.

HH-type estimator

Another unbiased estimator[1] can be formed by modifying the HH estimator making use of information of all n networks (which may not be distinct) generated by each person of the initial sample and their sizes. The modified HH type of estimator is

$$\hat{Y}_{\text{HH}(t)} = \frac{N_t}{n} \sum_{k=1}^{n} \frac{y^*_{k(t)}}{m_{k(t)}}. \tag{6}$$

Here, the total number of networks is taken to be n instead the number of distinct networks as taken in $\hat{Y}_{\text{HT}(t)}$. The network size, however, may vary from sample (initial) to sample. If any person from the initial sample is COVID-19-positive, then $y^*_{k(t)}/m_{k(t)}$ of the network induced by that person shall be unity, implying that $\hat{Y}_{\text{HH}(t)}$ becomes N_t if each person in the initial sample is positive. Hence, drawing the initial sample plays an important role for this particular type of estimator. However, this is not true when the parameters other than total counts are estimated.

The variance of $\hat{Y}_{\text{HH}(t)}$ is

$$\text{Var}(\hat{Y}_{\text{HH}(t)}) = \frac{N_t(N_t - n)}{n} S^2_{Bt}, \tag{7}$$

where S^2_{Bt} is between network variance at time t and is given by

$$S^2_{Bt} = \frac{1}{n-1} \sum_{k=1}^{n} \left(\frac{y^*_{k(t)}}{m_{k(t)}} - \bar{Y}^*_{k(t)}\right)^2, \quad \bar{Y}^*_{k(t)} = \frac{1}{n} \sum_{k=1}^{n} \frac{y^*_{k(t)}}{m_{k(t)}}.$$

Best linear unbiased estimator using small-domain estimators

In actual practice, the above mentioned two estimators might be more convenient for estimating the COVID-19-related parameters at a small domain level like districts or states due to their independent management of the disease according to available resources, geographical locations, fixing population size, etc. For the estimation of larger domains, viz. at the country level, the small domain estimates may be used. For this, the concept of best linear unbiased estimator (BLUE) is used considering that the small-level domain estimates are independent of each other and together they constitute the larger domain for which BLUE is to be obtained. In order to find BLUE for a large domain, the following theorem is used[4].

Theorem 1: Let $T_{1(t)}, T_{2(t)}, ..., T_{p(t)}$ be p independent estimators corresponding to the p small domains at time t with respective variances $\sigma^2_{1(t)}, \sigma^2_{2(t)}, ..., \sigma^2_{p(t)}$. Suppose the population of these p domains at the same time t are $N_{1(t)}, N_{2(t)}, ..., N_{p(t)}$ respectively, with $N_{1(t)} + N_{2(t)} + ... + N_{p(t)} = N_t$. The linear combination of estimators is

$$\hat{Y}_{B(t)} = \frac{N_{1(t)}}{N_t} T_{1(t)} + \frac{N_{2(t)}}{N_t} T_{2(t)} + ... + \frac{N_{p(t)}}{N_t} T_{p(t)}, \tag{8}$$

$\hat{Y}_{B(t)}$ has the smallest variance and is given by

$$\text{Var}(\hat{Y}_{B(t)}) = \frac{1}{\frac{1}{\sigma^2_{1(t)}} + \frac{1}{\sigma^2_{2(t)}} + ... + \frac{1}{\sigma^2_{p(t)}}}, \tag{9}$$

Since each of the $T_{i(t)}$'s, $i = 1, 2, ... , p$ are unbiased and $\hat{Y}_{B(t)}$ has the smallest variance, therefore $\hat{Y}_{B(t)}$ is considered as BLUE.

Empirical examples

Two empirical examples pertaining to Uttarakhand and Kerala are given here.

Uttarakhand

In Uttarakhand, the first case was detected on 15 March 2020, in a batch of 45 trainees who had come back from Spain, a COVID-19 hotspot. Two more trainees tested positive after three days and the remaining went into quarantine for 14 days. No more positive cases were detected in this batch. As on 5 April 2020, six persons tested positive from the same community in Dehradun (DD), and there were five districts in the state with at least one positive person as on this date (Table 1).

In that situation, actual survey was difficult due to lockdown and its extension from time to time, curfew in various places, sealing of corona hotspots, etc. However, using the data available from the Ministry of Health, Government of Uttarakhand (health.uk.gov.in), the following assumptions were made for computation purpose:

(i) The initial sample size of each district is same, i.e. $n = 50$.

(ii) The population size is taken as the counts of all those persons whose samples were sent for testing, persons who were in hospital quarantine, institutional isolation, etc. (health.uk.gov.in). This population size was rounded-off to the nearest value and a minimum of 100 (Table 2). The population size will definitely be different from the actual population of each district and the sampling is not actually done by simple random sampling.

(iii) The computations are based upon data of 5 April 2020 ($t = 0$).

(iv) With assumption (i) above, only the networks of DD and Udham Singh Nagar (US) districts were captured in the final sample having at least one COVID-19-positive person. The networks of remaining three districts having COVID-19 cases were left in the initial sample (Table 1). However, the possibility of inclusion of networks from these three districts and other COVID-19-negative districts still exists in the near future.

(v) While computing the estimates, only two distinct networks (trainees of size 3 and other communities of size 6 as mentioned above) were generated by the initial sample in DD. In US, all four COVID-19-positive persons formed a network of size 4, as they were considered from the same community.

With assumptions (i)–(v) above, the main aim of this example is to solve the estimation problem of districts and Uttarakhand through the BLUE of the district-level HT and HH estimates. The calculation of intersection probability, estimator and variances of both the districts and combined estimates of the State for 5 April 2020 ($t = 0$) is given below.

Dehradun: As there are two networks $A_{1(0)}$ and $A_{2(0)}$ (the smallest network denoted by $A_{1(0)}$ followed by $A_{2(0)}$, and so on) having at least one positive person with network size 3 and 6 respectively, the values of $\alpha_{1(0)}$ and $\alpha_{2(0)}$ for these two networks are

$$\alpha_{1(0)} = 1 - \left[\frac{\binom{500-3}{50}}{\binom{500}{50}} \right] = 0.2715 \text{ and}$$

$$\alpha_{2(0)} = 1 - \left[\frac{\binom{500-6}{50}}{\binom{500}{50}} \right] = 0.4703.$$

Table 1. District- and date-wise cumulative positive persons in Uttarakhand, India

	District												
Date	DD	NA	AL	HA	US	PG	PI	CL	BA	CH	TG	RP	UT
15 March	1	0	0	0	0	0	0	0	0	0	0	0	0
18 March	3	0	0	0	0	0	0	0	0	0	0	0	0
23 March	4	0	0	0	0	0	0	0	0	0	0	0	0
25 March	4	0	0	0	0	1	0	0	0	0	0	0	0
28 March	5	0	0	0	0	1	0	0	0	0	0	0	0
29 March	6	0	0	0	0	1	0	0	0	0	0	0	0
2 April	6	0	0	0	3	1	0	0	0	0	0	0	0
3 April	11	0	0	0	4	1	0	0	0	0	0	0	0
4 April	11	5	0	1	4	1	0	0	0	0	0	0	0
5 April	14	6	0	1	4	1	0	0	0	0	0	0	0
6 April	18	6	1	1	4	1	0	0	0	0	0	0	0
8 April	18	8	1	3	4	1	0	0	0	0	0	0	0
14 April	18	8	1	5	4	1	0	0	0	0	0	0	0
17 April	20	9	1	5	4	2	0	0	0	0	0	0	0
18 April	20	9	1	7	4	2	0	0	0	0	0	0	0

DD, Dehradun; NA, Nainital; AL, Almora; HA, Haridwar; US, Udham Singh Nagar; PG, Pauri Garhwal; PI, Pithoragarh; CL, Chamoli; BA, Bageshwar; CH, Champawat; TG, Tehri Garhwal; RP, Rudraprayag, UT, Uttarkashi.

Dates for which no positive cases were observed have been excluded from the table.

Table 2. District- and date-wise population size considered for the study of Uttarakhand

	District													
Time (*t*)	DD	NA	AL	HA	US	PG	PI	CL	BA	CH	TG	RP	UT	Total (N_t)
0	500	200	100	200	100	100	100	100	100	100	100	100	100	1,900
1	500	200	100	300	100	100	100	100	100	100	100	100	100	2,000
2	600	300	100	300	200	100	100	100	100	100	100	100	100	2,300
3	800	400	100	500	200	100	100	100	100	100	100	100	100	2,800
4	1,000	500	100	700	300	100	100	100	100	100	100	100	200	3,500
5	1,300	600	200	1,000	500	300	200	200	200	200	200	200	300	5,400

The HT-type estimator is

$$\hat{Y}_{\text{HT}(0)} = \frac{3}{0.2715} + \frac{6}{0.4703} = 23.80.$$

In order to compute $\hat{\text{Var}}(\hat{Y}_{\text{HT}(0)})$, we have $\alpha_{11(0)} = \alpha_{1(0)}$, $\alpha_{22(0)} = \alpha_{2(0)}$ and

$$\alpha_{12(0)} = \alpha_{21(0)}$$

$$= 1 - \frac{\left[\binom{500-3}{50} + \binom{500-6}{50} - \binom{500-3-6}{50}\right]}{\binom{500}{50}} = 0.1261.$$

Using eq. (5) $\hat{\text{Var}}(\hat{Y}_{\text{HT}(0)}) = 171.60.$

The HH-type estimators is

$$\hat{Y}_{\text{HH}(0)} = \frac{500}{50} \times 2 = 20.$$

We have

$$S_{B0}^2 = \frac{(1-0.04)^2 + (1-0.04)^2}{49} = 0.0376.$$

Therefore

$$\text{Var}(\hat{Y}_{\text{HH}(0)}) = \frac{500 \times 450}{50} 0.0376 = 169.20.$$

In order to compute the variances under simple random sampling, the binomial distribution with parameters

$$N_0 = 500,\ p_0 = \frac{14}{500},\ q_0 = \frac{486}{500}$$

is used.

The variance of binomial variate is given by $N_0 p_0 q_0 = 13.61$. Therefore, the variance of population total under SRS shall be

$$\text{Var}(\hat{Y}_{\text{SRS}}) = 500 \times 13.61 = 6804.$$

The relative precision (RP) of HT- and HH-type estimators with SRS is the obtained (Table 3).

Udham Singh Nagar: The value of $\alpha_{1(0)}$ for the sole network $A_{1(0)}$ of size 4 is given by

$$\alpha_{1(0)} = 1 - \left[\frac{\binom{100-4}{50}}{\binom{100}{50}} \right] = 0.9412.$$

This gives

$$\hat{Y}_{\mathrm{HT}(0)} = \frac{4}{0.9412} = 4.25,$$

with

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{Y}_{\mathrm{HT}(0)}) = y_{1(0)}^{*2} \left(\frac{1-\alpha_{1(0)}}{\alpha_{1(0)}^2} \right) = 1.06.$$

Using eq. (6)

$$\hat{Y}_{\mathrm{HH}(0)} = \frac{100}{50} = 2.$$

Further,

$$S_{B0}^2 = \frac{(1-0.02)^2}{49} = 0.0196.$$

This gives

$$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(0)}) = \frac{100\times 50}{50} 0.0196 = 1.96.$$

The variance of population total under SRS using the binomial distribution with

$$N_0 = 100,\ p_0 = \frac{4}{100},\ q_0 = \frac{96}{100},$$

is given by

$$\mathrm{Var}(\hat{Y}_{\mathrm{SRS}}) = 384.$$

Table 3 shows the values of RP.

BLUE for Uttarakhand state: While finding the estimators and their variances by combining the district-level estimates, theorem 1 has been used. On 5 April 2020, the estimators are available for two districts only. The values of coefficients $N_{i(0)}/N_0$'s, $i = 1, 2, ..., 13$ are simply the proportions of population size of the *i*th district. With the different values of proportion of population size, BLUE for HT-type estimator is given as

$$\hat{Y}_{\mathrm{BHT}(0)} = \frac{500}{1900} 23.8 + \frac{100}{1900} 4.25 = 6.49,$$

with

$$\mathrm{Var}(\hat{Y}_{\mathrm{BHT}(0)}) = \frac{1}{\frac{1}{171.60} + \frac{1}{1.06}} = 1.05.$$

Similarly $\hat{Y}_{\mathrm{BHH}(0)} = 5.37$

$$\mathrm{Var}(\hat{Y}_{\mathrm{BHH}(0)}) = \frac{1}{\frac{1}{169.20} + \frac{1}{1.96}} = 1.94.$$

The variance of population total of Uttarakhand under SRS is computed using theorem 1 and is given as

$$\mathrm{Var}(\hat{Y}_{\mathrm{BSRS}(t=0)}) = \frac{1}{\frac{1}{6804} + \frac{1}{384}} = 363.49.$$

Table 3 gives the values of RP.

Dynamic variation of estimates from 5 to 18 April: With the same procedure for district and state computations as given earlier, the intersection probabilities, estimators and their variances for five other dates, namely 5 April ($t = 0$), 6 April ($t = 1$), 8 April ($t = 2$), 14 April ($t = 3$), 17 April ($t = 4$) and 18 April 2020 ($t = 5$) were calculated (Table 3). These are the dates in which at least one case was detected from 5 to 18 April 2020. They provide the variation of estimates as seen in actual practice. While determining these variations, the following additional assumptions are considered.

(i) In DD, at $t = 1$ one more network ($A_{3(1)}$) of size 3 was added. At $t = 4$, the size of this network $A_{3(4)}$ was increased by one. The size of the other two networks was the same up to $t = 5$.

(ii) There was no change in network size and the number of networks in US.

(iii) There was inclusion of one network of size 2 in Nainital (NT) district at $t = 2$ and the size of this network was increased by 1 at $t = 4$.

Kerala

Another example, Kerala, is taken in which the first positive person was detected on 30 January 2020. The data are taken from the official website: https://dashboard.kerala.gov.in/ and are for the same period $t = 0$ to $t = 5$ (Table 4). Besides the assumptions already mentioned earlier, the following are considered for computations:

(i) Two distinct networks (size 3 and 6 respectively) were generated by the initial sample at $t = 0$ in each of the two districts Thiruvananthapuram (TV) and Kasaragod (KG) of sizes (3, 6) and (3, 10) respectively.

(ii) Under variation in the number of networks and their size: (a) There was no change of network size and

Table 3. Dynamic variation of intersection probabilities, estimates and best linear unbiased estimator (BLUE) over time for the selected districts and Uttarakhand

		Time					
District	Coefficients/values	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
DD	$\alpha_{1(t)}$	0.2715	0.2715	0.2301	0.1762	0.1428	0.1111
	$\alpha_{2(t)}$	0.4703	0.4703	0.4081	0.3219	0.2655	0.2101
	$\alpha_{3(t)}$	–	0.2715	0.2301	0.1762	0.1858	0.1454
	$\alpha_{12(t)}$	0.1261	0.1261	0.0927	0.0559	0.0373	0.0229
	$\alpha_{13(t)}$	–	0.0726	0.0522	0.0305	0.0261	0.0159
	$\alpha_{23(t)}$	0.4703	0.1261	0.0927	0.0559	0.0485	0.0300
	$\hat{Y}_{\mathrm{HT}(t)}$	23.81	34.86	40.78	52.69	65.15	83.09
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	171.60	253.41	374.26	683.92	1085.31	1862.90
	$RP_{\mathrm{HT(t)}}$	39.65	34.24	27.99	20.58	18.06	13.74
	$\hat{Y}_{\mathrm{HH}(t)}$	20	30	36	48	60	78
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	169.27	243.44	357.05	649.18	1027.86	1758.18
	$RP_{\mathrm{HH}}(t)$	40.20	35.64	29.34	21.68	19.07	14.56
US	$\alpha_{1(t)}$	0.9412	0.9412	0.6868	0.6868	0.5197	0.3448
	$\hat{Y}_{\mathrm{HT}(t)}$	4.25	4.25	5.82	5.82	7.70	11.60
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	1.06	1.06	10.63	10.63	28.46	88.21
	$RP_{\mathrm{HT}}(t)$	362.26	362.26	73.75	73.75	41.60	22.49
	$\hat{Y}_{\mathrm{HH}(t)}$	2	2	4	4	6	10
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	1.96	1.96	11.76	11.76	29.40	88.20
	$RP_{\mathrm{HH}(t)}$	195.92	195.92	66.67	66.67	40.27	22.49
NA	$\alpha_{1(t)}$	–	–	0.3060	0.3308	0.2715	0.2301
	$\hat{Y}_{\mathrm{HT}(t)}$	–	–	6.54	6.05	11.05	13.04
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	–	–	29.65	24.47	88.97	130.85
	$RP_{\mathrm{HT}}(t)$	–	–	30.05	80.71	27.82	31.72
	$\hat{Y}_{\mathrm{HH}(t)}$	–	–	6	8	10	12
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	–	–	29.40	54.88	88.20	129.36
	$RP_{\mathrm{HH}(t)}$	–	–	30.31	35.99	28.06	32.09
BLUE (Uttarakhand)	$\hat{Y}_{\mathrm{BHT}(t)}$	6.49	8.93	12.00	16.33	20.85	22.53
	$\mathrm{Var}(\hat{Y}_{\mathrm{BHT}(t)})$	1.05	1.06	7.66	7.33	21.14	51.24
	$RP_{\mathrm{BHT}(t)}$	346.18	346.91	52.36	73.63	36.40	24.89
	$\hat{Y}_{\mathrm{BHH}(t)}$	5.37	7.60	10.52	15.14	19.09	21.04
	$\mathrm{Var}(\hat{Y}_{\mathrm{BHH}(t)})$	1.94	1.94	8.21	9.54	21.59	50.92
	$RP_{\mathrm{BHH}(t)}$	187.36	189.55	48.85	56.57	35.64	25.05

Table 4. District- and date-wise cumulative positive persons in Kerala

	District													
Time (*t*)	TV	MP	EK	KG	AL	KK	PK	KL	TS	PT	KN	KT	ID	WY
0	12	13	21	143	3	7	7	6	12	14	51	3	10	3
1	12	15	21	152	3	7	7	7	12	15	51	3	10	3
2	12	16	21	155	5	7	7	8	13	16	57	3	10	3
3	13	20	21	165	5	11	8	9	13	17	65	3	10	3
4	13	20	21	166	5	14	8	9	13	17	70	3	10	3
5	13	20	21	166	5	15	8	9	13	17	73	3	10	3

TV, Thiruvananthapuram; MP, Malappuram; EK, Ernakulam; KG, Kasaragod; AL, Alappuzha; KK, Kozhikode; PK, Palakkad; KL, Kollam; TS, Thrissur; PT, Pathanamthitta; KN, Kannur; KT, Kottayam; ID, Idukki; WY, Wayanad.

Table 5. District- and date-wise population size considered for Kerala

	District														
Time (t)	TV	MP	EK	KG	AL	KK	PK	KL	TS	PT	KN	KT	ID	WY	Total (N_t)
0	1,700	1,500	800	2,100	400	700	800	300	1,300	600	1,600	200	200	200	12,400
1	1,800	1,700	900	2,300	400	700	900	300	1,400	600	1,700	200	200	200	13,300
2	2,000	1,900	900	2,800	400	700	900	400	1,400	600	1,900	200	200	200	14,500
3	2,600	2,700	1,100	4,200	500	900	1,100	400	1,500	700	2,500	200	200	300	18,900
4	3,000	2,800	1,200	4,600	500	1,000	1,200	500	1,600	800	2,900	200	400	300	21,000
5	3,100	2,800	1,200	4,700	500	1,000	1,300	500	1,600	800	3,000	300	500	300	21,600

Table 6. Dynamic variation of intersection probabilities, estimates and BLUE over time for the selected districts and Kerala

		Time					
District	Coefficients/values	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
TV	$\alpha_{1(t)}$	0.0857	0.0811	0.0732	0.0567	0.0492	0.0476
	$\alpha_{2(t)}$	0.1642	0.1557	0.1411	0.1101	0.0960	0.0930
	$\alpha_{12(t)}$	0.0138	0.0124	0.0102	0.0061	0.0047	0.0043
	$\hat{Y}_{\mathrm{HT}(t)}$	71.54	75.53	83.51	107.46	123.42	127.50
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	2192.34	2460.69	3055.34	5160.49	6953.31	7390.03
	$RP_{\mathrm{HT}(t)}$	9.24	8.72	7.81	6.52	5.58	5.43
	$\hat{Y}_{\mathrm{HH}(t)}$	68	72	80	104	120	124
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	2110.28	2369.83	2934.07	4987.92	6658.09	7113.25
	$RP_{\mathrm{HH}(t)}$	9.60	9.05	8.13	6.74	5.83	5.64
KG	$\alpha_{1(t)}$	0.0698	0.0638	0.0526	0.0353	0.0323	0.0316
	$\alpha_{2(t)}$	0.2145	0.1977	0.1651	0.1130	0.1036	0.1015
	$\alpha_{3(t)}$	–	0.0638	0.0526	0.0353	0.1789	0.1754
	$\alpha_{12(t)}$	0.0147	0.0124	0.0085	0.0039	0.0033	0.0031
	$\alpha_{13(t)}$	–	–0.0700	–0.0598	–0.0424	0.0057	0.0054
	$\alpha_{23(t)}$	–	0.0124	0.0085	0.0039	0.0182	0.0175
	$\hat{Y}_{\mathrm{HT}(t)}$	89.59	144.62	174.54	258.47	290.05	296.13
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	3353.55	10676.68	15758.67	35270.57	24169.28	25025.87
	$RP_{\mathrm{HT}(t)}$	83.45	30.58	26.02	18.88	30.45	30.07
	$\hat{Y}_{\mathrm{HH}(t)}$	84	138	168	252	276	282
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	3238.77	5599.14	8331.09	18858.55	22645.41	23646.22
	$RP_{\mathrm{HH}(t)}$	86.41	58.31	49.21	35.30	32.50	31.83
PT	$\alpha_{1(t)}$	–	–	0.3538	0.4061	0.3646	0.3646
	$\hat{Y}_{\mathrm{HT}(t)}$	–	–	14.13	12.31	13.71	19.20
	$\mathrm{Var}(\hat{Y}_{\mathrm{HT}(t)})$	–	–	129.10	90.02	119.48	234.19
	$RP_{\mathrm{HT}(t)}$	–	–	72.38	128.98	111.41	56.84
	$\hat{Y}_{\mathrm{HH}(t)}$	–	–	12	14	16	16
	$\mathrm{Var}(\hat{Y}_{\mathrm{HH}(t)})$	–	–	129.36	178.36	235.20	235.20
	$RP_{\mathrm{HH}(t)}$	–	–	72.23	65.10	56.59	56.59
BLUE (Kerala)	$\hat{Y}_{\mathrm{BHT}(t)}$	24.98	35.23	45.81	72.68	81.69	83.45
	$\mathrm{Var}(\hat{Y}_{\mathrm{BHT}(t)})$	1325.69	1999.79	122.90	88.26	116.89	224.96
	$RP_{\mathrm{BHT}(t)}$	14.25	10.07	53.75	96.55	83.68	43.85
	$\hat{Y}_{\mathrm{BHH}(t)}$	23.55	33.61	43.97	70.83	78.21	79.75
	$\mathrm{Var}(\hat{Y}_{\mathrm{BHH}(t)})$	1277.74	1665.08	122.08	170.64	224.92	225.50
	$RP_{\mathrm{BHH}(t)}$	14.78	12.09	54.11	49.93	43.49	43.75

number of networks in TV. (b) At $t = 1$, one more network of size 3 was added in KG. At $t = 4$, the size of this network increased by 15. (c) There was inclusion of one network of size 5 in Pathanamthitta (PT) at $t = 2$ and the size of this network increased by 2 at $t = 5$.

Table 5 shows date and district-wise population size for the computations. The results were obtained (Table 6) using the same procedure as mentioned earlier.

Discussion on empirical examples

Tables 3 and 6 suggest that as time increases, there might be variation (either in increasing or decreasing order) in the number of networks, their size estimates, intersection probabilities and RP with respect to conventional SRS. In spite of both estimators providing similar results, HT-type of estimator may be opted in general as it depends upon the intersection probability and may not provide overestimates in compared to the HH-type estimator. When there are more cases in the initial sample, HH-type of estimator may produce overestimates. In all cases, RP > 1. In these particular examples, the estimates are closer to the reported values to some extent. It may, however, vary from example to example. The estimates should be nearer to the true value if (i) there is 100% testing for obtaining the true value, (ii) initial sample is drawn on a purely random basis from the hotspot areas and (iii) the neighbourhood is properly defined for adapting from them in the final sample.

Conclusion

The design proposed here may help to find precise estimates of COVID-19 cases and hotspots in a region using HT and HH-type estimators. Both estimators depend upon the population size. It is suggested to take the population size as the aggregate of persons directly or indirectly related to those of the initial sample and their neighbourhoods for similar studies. Considering the total persons of the domain might be appropriate when COVID-19 cases are uniformly spread over the domain. HT provides information about possibility of any person being COVID-19-positive and belonging to the network A_i over time.

The drawing of the initial sample plays a crucial role in predicting the number of affected persons over time. In the present case, the rate of increase is much higher in all the countries. Accordingly, the sample should be drawn in such a way that the same rate is maintained in the initial sample, especially for those who were quarantined while choosing the persons in the final sample. Therefore, the network size should be increased in the same way. In addition, the final sample might be a useful source for developing mathematical models for prediction. In the current situation, many countries face a problem of testing all the affected or likely to be affected persons. Those affected by COVID-19 are not captured due to the non-availability of resources, or other reasons. In such cases the adaptive design might be a helpful tool in tracing or finding the distribution of COVID-19-affected persons.

The Government of India categorized the districts/ geographical areas into three different zones, namely red, orange and green, depending upon the number of people infected. The ACS networks also give the same information. The pre-specified number of persons found to be COVID-19-positive in the first phase of ACS may be considered as the red category. The pre-specified number of persons found positive within the quarantine-period may be considered as orange category and the edge persons, whether found positive or not, may be considered as green category. Similar information might be generated through the networks or clusters of ACS, which is recommended in this article.

1. Thompson, S. K., Adaptive cluster sampling. *J. Am. Stat. Assoc.*, 1990, **85**, 1050–1059.
2. Thompson, S. K., Adaptive cluster sampling based on order statistics. *Environmetrics*, 1996, **7**(2), 123–133.
3. Chandra, G., Tiwari, N. and Nautiyal, R., Two stage adaptive cluster sampling based on ordered statistics. *Metodoloski Zv.*, 2019, **16**(1), 43–60.
4. Kaur, A., Patil, G. P. and Taillie, C., Optimal allocation for symmetric distributions in ranked sampling. *Ann. Inst. Stat. Math.*, 2000, **52**(2), 250.

Received 23 May 2020; revised accepted 23 December 2020

doi: 10.18520/cs/v120/i7/1202-1210