

# Estimation in Overlapping Clusters with Unknown Population Size

D.S. TRACY and S.S. OSAHAN<sup>1</sup>

## ABSTRACT

Two sampling strategies for estimation of population mean in overlapping clusters with known population size have been proposed by Singh (1988). In this paper, ratio estimators under these two strategies are studied assuming the actual population size to be unknown, which is the more realistic situation in sample surveys. The sampling efficiencies of the two strategies are compared and a numerical illustration is provided.

KEY WORDS: Overlapping clusters; Clustering before sampling; Mean square error; Relative efficiency.

## 1. INTRODUCTION

In cluster sampling, clusters are formed either before selecting the sample (CBS) or after selecting the sample (CAS). In both cases, clusters may be overlapping or non-overlapping. For non-overlapping clusters, much work by several researchers is available in the literature. However, there are many practical sampling situations where one gets overlapping clusters. For example, overlapping clusters may exist in some regional epidemiological survey for a contagious disease like mycobacterium tuberculosis (T.B.), becoming very prevalent with the spread of AIDS (Gifford-Jones 1993). Clusters here may be formed around infected individuals or closely associated individuals who are more vulnerable to the same type of infection. A similar situation may exist in an ecological survey where clusters are formed around the factories burning coal and emitting polyaromatic hydrocarbons (PAH's) which are potent cancer causing compounds. Clusters are formed on the basis of the intensity of such gases, and surveys may be required in order to control air pollution which causes lung diseases like bronchitis. For overlapping clusters, one can refer to the limited work done by Goel and Singh (1977), Agarwal and Singh (1982) and Amdekar (1985). But the methodologies developed by them suffer from one limitation or the other.

Recently, Singh (1988) has developed a very simple estimator for a population mean using two sampling strategies in the CBS system assuming known population size. In the first strategy, clusters are selected with equal probabilities, whereas in the second case selection probabilities are taken proportional to cluster size. The elements within the clusters are selected with equal probability in both the cases. But it is unrealistic to assume that the actual population size is known. If it is the case, then all the duplicates in the population are known *a priori*, and one

could easily remove them to increase the efficiency of the sampling design. Hence, the estimators of the population mean studied by Singh (1988) need an improvement in order to be practicable, as they depend on the actual population size. This limitation in the methodology has motivated the present work.

We propose two sampling strategies in the CBS system with simple ratio estimators for the population mean, which do not depend on the actual population size. As in Singh (1988), an equal probability with replacement sampling scheme is used for selecting the clusters in the first strategy, whereas in the second, an unequal probability sampling scheme is used. The elements within the clusters are selected with an equal probability without replacement sampling scheme in both strategies.

The population of  $N$  units under consideration is expressible in the form of  $K$  overlapping clusters with  $N_i$  units in the  $i$ -th cluster and  $\sum_{i=1}^K N_i = M \geq N$ , the unknown actual population size, (equality holds only for non-overlapping clusters). A population unit may be included in more than one cluster. Let  $y$  be the characteristic of interest and let the population mean be  $\bar{Y}$ .

Define

$$Z_{ij} = Y_{ij}/F_{ij}, \quad W_{ij} = 1/F_{ij}; \quad i = 1, 2, \dots, K,$$
$$\text{and } j = 1, 2, \dots, N_i$$

where  $Y_{ij}$  is the value of  $y$  for the  $j$ -th unit in the  $i$ -th cluster and  $F_{ij}$  its frequency of occurring in  $K$  clusters.

When clusterwise data on units are available on the computer, the values of these frequencies for overlapping clusters may be easily available. As for the example considered earlier in epidemiology, suppose we have data available for households or individuals along with their identification labels like house numbers or social insurance

<sup>1</sup> D.S. Tracy and S.S. Osahan, Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario N9B 3P4.

numbers/health card numbers on the computer. Then, by giving a simple command to the computer, a researcher can easily extract information about the repetition of a certain unit from its label in different clusters. Also, in case we have a map of the overlapping clusters and the criterion for forming clusters does not allow the elimination of duplicacy of units in the different clusters, the values of such frequencies may be known.

The two strategies are discussed in section 2 and their efficiencies are compared in section 3.

## 2. THE TWO STRATEGIES

The two proposed strategies are discussed in Sections 2.1 and 2.2. Their comparison is undertaken in Section 3.

### 2.1 Strategy A

This strategy consists of the following steps:

- Select  $k$  clusters out of  $K$  by simple random sampling with replacement (SRSWR).
- From the  $i$ -th selected cluster of size  $N_i$  ( $i = 1, \dots, K$ ), select  $n_i$  elementary units by simple random sampling without replacement (SRSWOR).

**Theorem 1.** The ratio estimator under SRS

$$\bar{z}_{RS} = \hat{Y}_{RS} / \hat{N}_{RS} = \frac{K}{k} \sum_{i=1}^k N_i \bar{z}_i \Big/ \frac{K}{k} \sum_{i=1}^k N_i \bar{w}_i \quad (1)$$

has relative bias, to the first order of approximation,

$$RB(\bar{z}_{RS}) \doteq \frac{K}{k} \left[ \left( \frac{\sigma_{bw}^2}{N^2} - \frac{\sigma_{bzw}}{NY} \right) K + \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) \left( \frac{S_{iw}^2}{N^2} - \frac{S_{izw}}{NY} \right) \right] \quad (2)$$

where

$$\sigma_{bzw} = \sum_{i=1}^K (N_i \bar{Z}_i - Y/K) (N_i \bar{W}_i - N/K) / K$$

$$S_{izw} = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i) (W_{ij} - \bar{W}_i) / (N_i - 1),$$

$$\bar{Z}_i = \sum_{j=1}^{N_i} Z_{ij} / N_i \quad \text{and} \quad \bar{z}_i = \sum_{j=1}^{n_i} z_{ij} / n_i,$$

and  $\sigma_{bw}^2$ ,  $S_{iw}^2$ ,  $\bar{W}_i$  and  $\bar{w}_i$  are the expressions of  $\sigma_{bzw}$ ,  $S_{izw}$ ,  $\bar{Z}_i$  and  $\bar{z}_i$  respectively, with  $z$  replaced by  $w$  and  $Y$  replaced by  $N$ .

**Proof.** Following a standard result, the relative bias of the estimator  $\bar{z}_{RS}$ , to the first order of approximation, is

$$RB(\bar{z}_{RS}) \doteq [V(\hat{N}_{RS})/N^2] - \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS})/YN. \quad (3)$$

Let  $E_2$  and  $V_2$  denote the conditional expectation and variance for a given sample of clusters and  $E_1$  and  $V_1$  the expectation and variance over all such samples. Then, we have

$$\begin{aligned} V(\hat{N}_{RS}) &= V_1 E_2(\hat{N}_{RS}) + E_1 V_2(\hat{N}_{RS}) \\ &= V_1 \left[ \frac{K}{k} \sum_{i=1}^k N_i E_2(\bar{w}_i) \right] \\ &\quad + E_1 \left[ \frac{K^2}{k^2} \sum_{i=1}^k N_i^2 V_2(\bar{w}_i) \right] \\ &= V_1 \left( \frac{K}{k} \sum_{i=1}^k N_i \bar{W}_i \right) \\ &\quad + E_1 \left[ \frac{K^2}{k^2} \sum_{i=1}^k N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2 \right] \\ &= \frac{K^2}{k} \sigma_{bw}^2 + \frac{K}{k} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2. \quad (4) \end{aligned}$$

Similarly, we have

$$\begin{aligned} \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS}) &= \frac{K^2}{k} \sigma_{bzw} \\ &\quad + \frac{K}{k} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw}. \quad (5) \end{aligned}$$

By substituting (4) and (5) in (3), we obtain (2), which completes the proof of the theorem.

**Theorem 2.** The mean square error (MSE) of the estimator  $\bar{z}_{RS}$ , to the first order of approximation, is

$$\begin{aligned} \text{MSE}(\bar{z}_{RS}) &\doteq \\ &\frac{K}{kN^2} \sum_{i=1}^K N_i^2 \left[ (\bar{Z}_i - \bar{Y} \bar{W}_i)^2 + \left( \frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right] \quad (6) \end{aligned}$$

where  $D_i^2 = S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2 S_{iw}^2$ , and  $S_{iz}^2 = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)^2 / (N_i - 1)$ .

**Proof.** To the first order of approximation, we have

$$\text{MSE}(\bar{z}_{RS}) \doteq [V(\hat{Y}_{RS}) - 2\bar{Y} \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS}) + \bar{Y}^2 V(\hat{N}_{RS})] / N^2. \quad (7)$$

The expression for  $V(\hat{Y}_{RS})$  may be written, following (4), as

$$V(\hat{Y}_{RS}) = \frac{K^2}{k} \sigma_{bz}^2 + \frac{K}{k} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2 \quad (8)$$

where  $\sigma_{bz}^2 = \sum_{i=1}^K (N_i \bar{Z}_i - Y/K)^2 / K$ .

By substituting (4), (5) and (8) in (7), we obtain upon simplification

$$\begin{aligned} \text{MSE}(\bar{z}_{RS}) &\doteq \frac{K^2}{kN^2} (\sigma_{bz}^2 - 2\bar{Y}\sigma_{bz w} + \bar{Y}^2 \sigma_{bw}^2) \\ &+ \frac{K}{kN^2} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) (S_{iz}^2 - 2\bar{Y} S_{iz w} + \bar{Y}^2 S_{iw}^2). \quad (9) \end{aligned}$$

Substitution of the expressions for  $\sigma_{bz}^2$ ,  $\sigma_{bz w}$  and  $\sigma_{bw}^2$  into (9) and simplification yields (6). Now, we provide an estimator of  $\text{MSE}(\bar{z}_{RS})$  below.

**Theorem 3.** A consistent estimator of  $\text{MSE}(\bar{z}_{RS})$ , to the first order of approximation, is given by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2} \cdot \frac{1}{k-1} \sum_{i=1}^k N_i^2 (\bar{z}_i - \bar{z}_{RS} \bar{w}_i)^2. \quad (10)$$

**Proof.** We note that the first-stage sampling is done with SRSWR sampling scheme and the random variables  $N_i \bar{z}_i$  and  $N_i \bar{w}_i$  in the ratio estimator are independently and identically distributed. Hence, the mean square error of  $\bar{z}_{RS}$  can be estimated using the well-known result that a variance estimator for a multi-stage design can consider the first stage only (see Särndal, Swensson and Wretman, 1992, Results 2.9.1 and 4.5.1).

From (9), an unbiased estimator of

$$\sigma_{bz}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2$$

can be written as

$$s_{bz}^2 = \frac{1}{k-1} \sum_{i=1}^k \left( N_i \bar{z}_i - \sum_{i=1}^k N_i \bar{z}_i / k \right)^2, \quad (11)$$

and an unbiased estimator of

$$\sigma_{bz w} + \frac{1}{K} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz w}$$

is

$$\begin{aligned} s_{bz w} &= \frac{1}{k-1} \sum_{i=1}^k \left( N_i \bar{z}_i - \sum_{i=1}^k N_i \bar{z}_i / k \right) \\ &\times \left( N_i \bar{w}_i - \sum_{i=1}^k N_i \bar{w}_i / k \right). \quad (12) \end{aligned}$$

Similarly, an independent estimator of

$$\sigma_{bw}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2$$

is  $s_{bw}^2$ , defined parallel to (11).

Using these results, one can easily show that a consistent estimator of  $\text{MSE}(\bar{z}_{RS})$  given in (6) is provided by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2} (s_{bz}^2 - 2\bar{z}_{RS} s_{bz w} + \bar{z}_{RS}^2 s_{bw}^2),$$

which can be written as (10).

## 2.2 Strategy B

This strategy consists of the following steps:

- (a) Select  $k$  clusters out of  $K$  by probability proportional to size with replacement (PPSWR) sampling with selection probabilities  $P_i = N_i/M$ ,  $i = 1, \dots, K$ .
- (b) Same as for strategy A.

**Theorem 4.** The ratio estimator under PPS sampling

$$\bar{z}_{RP} = \hat{Y}_{RP} / \hat{N}_{RP} = \frac{M}{k} \sum_{i=1}^k \bar{z}_i \Big/ \frac{M}{k} \sum_{i=1}^k \bar{w}_i \quad (13)$$

has relative bias, to the first order of approximation,

$$\begin{aligned} \text{RB}(\bar{z}_{RP}) &\doteq \frac{M^2}{k} \left[ \left( \frac{\sigma_{bw'}^2}{N^2} - \frac{\sigma_{bz w'}}{YN} \right) \right. \\ &\left. + \sum_{i=1}^K \frac{N_i}{M} \left( \frac{1}{n_i} - \frac{1}{N_i} \right) \left( \frac{S_{iw}^2}{N^2} - \frac{S_{iz w}}{YN} \right) \right] \quad (14) \end{aligned}$$

where

$$\sigma_{bz w'} = \sum_{i=1}^K (\bar{Z}_i - Y/M) (\bar{W}_i - N/M) (N_i/M)$$

and  $\sigma_{bw'}$  is the expression of  $\sigma_{bz w'}$  with  $z$  replaced by  $w$  and  $Y$  replaced by  $N$ .

**Proof.** Using a standard result, the approximate relative bias, to the first order of approximation, is

$$RB(\bar{z}_{RP}) \doteq [V(\hat{N}_{RP})/N^2] - \text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP})/YN. \quad (15)$$

We have

$$\begin{aligned} V(\hat{N}_{RP}) &= V_1 E_2(\hat{N}_{RP}) + E_1 V_2(\hat{N}_{RP}) \\ &= M^2 \left[ V_1 \frac{1}{k} \sum_{i=1}^k E_2(\bar{w}_i) + E_1 \frac{1}{k^2} \sum_{i=1}^k V_2(\bar{w}_i) \right] \\ &= M^2 \left[ V_1 \frac{1}{k} \sum_{i=1}^k \bar{W}_i + E_1 \frac{1}{k^2} \sum_{i=1}^k \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2 \right] \\ &= \frac{M^2}{k} \left[ \sigma_{bw'}^2 + \sum_{i=1}^K \frac{N_i}{M} \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2 \right]. \quad (16) \end{aligned}$$

Similarly, one can write

$$\text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) = \frac{M^2}{k} \left[ \sigma_{bz'w'} + \sum_{i=1}^K \frac{N_i}{M} \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw} \right]. \quad (17)$$

Substituting (16) and (17) in (15), we obtain (14).

**Theorem 5.** The MSE of the estimator  $\bar{z}_{RP}$ , to the first order of approximation, is

$$\begin{aligned} \text{MSE}(\bar{z}_{RP}) &\doteq \frac{M}{kN^2} \sum_{i=1}^K N_i \\ &\times \left[ (\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left( \frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right]. \quad (18) \end{aligned}$$

**Proof.** We write, to the first order of approximation,

$$\begin{aligned} \text{MSE}(\bar{z}_{RP}) &\doteq [V(\hat{Y}_{RP}) - 2\bar{Y} \text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) \\ &\quad + \bar{Y}^2 V(\hat{N}_{RP})]/N^2. \quad (19) \end{aligned}$$

Also, from Theorem 2.5 of Singh (1988), we have by analogy

$$V(\hat{Y}_{RP}) = \frac{M^2}{k} \sigma_{bz'}^2 + \frac{M^2}{k} \sum_{i=1}^K \frac{N_i}{M} \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2, \quad (20)$$

where  $\sigma_{bz'}^2 = \sum_{i=1}^K (N_i/M) (\bar{Z}_i - Y/M)^2$ . On substituting (16), (17) and (20) in (19) and simplifying, we obtain (18).

**Theorem 6.** A consistent estimator of  $\text{MSE}(\bar{z}_{RP})$ , to the first order of approximation, is

$$\widehat{\text{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2} \cdot \frac{1}{k(k-1)} \sum_{i=1}^k (\bar{z}_i - \bar{z}_{RP} \bar{w}_i)^2. \quad (21)$$

**Proof.** As the first-stage units are selected with PPSWR, the justification given in the proof of theorem 3 applies here, as well.

From (20), using Results 2.9.1 and 4.5.1 of Särndal, Swensson and Wretman (1992), an unbiased estimator of

$$\sigma_{bz'}^2 + \sum_{i=1}^K \frac{N_i}{M} \left( \frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2$$

can be written as

$$s_{bz'}^2 = \frac{1}{k-1} \sum_{i=1}^k \left( \bar{z}_i - \sum_{i=1}^k \bar{z}_i/k \right)^2. \quad (22)$$

Similarly, defining  $s_{bz'w'}$  and  $s_{bw'}^2$ , one can show that

$$\widehat{\text{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2 k} (s_{bz'}^2 - 2\bar{z}_{RP} s_{bz'w'} + \bar{z}_{RP}^2 s_{bw'}^2),$$

which can be written as (21).

### 3. EFFICIENCY COMPARISON

The efficiencies of the estimators are compared below under the two strategies.

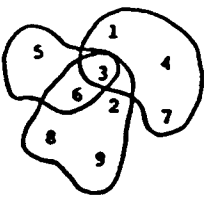
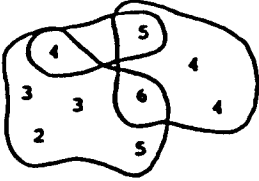
**Remark.** The estimator  $\bar{z}_{RP}$  under strategy B is expected to be more efficient than the estimator  $\bar{z}_{RS}$  under strategy A.

We provide a justification. From (6) and (18), we obtain

$$\begin{aligned} \text{MSE}(\bar{z}_{RS}) - \text{MSE}(\bar{z}_{RP}) &\doteq \frac{M}{kN^2} \sum_{i=1}^K N_i \\ &\times \left[ (\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left( \frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right] \left( \frac{KN_i}{M} - 1 \right). \end{aligned}$$

As the cluster size  $N_i$  increases, the factor  $(KN_i/M - 1)$  will also increase. The other factor of the term under summation is  $N_i [(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i) D_i^2]$ , which represents the contribution due to variability in  $z$  and  $w$  present in the  $i$ -th cluster (without the constant  $M/kN^2$ ) towards  $\text{MSE}(\bar{z}_{RP})$  in (18). As cluster size  $N_i$  increases, the contribution of the  $i$ -th cluster towards  $\text{MSE}(\bar{z}_{RP})$  is also expected to increase. This makes the covariance between these two factors positive. Hence, the estimator  $\bar{z}_{RP}$  is expected to have a smaller MSE than  $\bar{z}_{RS}$ .

**Table 1**  
Comparison of the Two Strategies for Two Small Populations

	Population No. 1			Population No. 2		
						
$N_i$	3	4	5	2	4	6
$n_i$	1	2	2	1	2	2
$Y_{ij}$	3,5,6	1,3,4,7	2,3,6,8,9	4,5	4,4,5,6	2,3,3,4,5,6
$F_{ij}$	3,1,2	1,3,1,1	1,3,2,1,1	2,2	1,1,2,2	1,1,1,2,1,2
$Z_{ij}$	1,5,3	1,1,4,7	2,1,3,8,9	2,2,5	4,4,2,5,3	2,3,3,2,5,3
$W_{ij}$	$\frac{1}{2}, 1, \frac{1}{2}$	$1, \frac{1}{3}, 1, 1$	$1, \frac{1}{3}, \frac{1}{2}, 1, 1$	$\frac{1}{2}, \frac{1}{2}$	$1, 1, \frac{1}{2}, \frac{1}{2}$	$1, 1, 1, \frac{1}{2}, 1, \frac{1}{2}$
$F$	1.38	10.16	18.12	.24	.77	2.94
MSE( $\bar{z}_{RS}$ )		2.09			0.45	
MSE( $\bar{z}_{RP}$ )		1.83			0.33	
R.E.		114.21			136.36	
R.B. ( $\bar{z}_{RS}$ )		-.0105			.0348	
R.B. ( $\bar{z}_{RP}$ )		-.0047			-.0037	

**Numerical Illustration.** Here the two proposed sampling strategies are applied to two small populations to shed light on the computations of  $F_{ij}$ ,  $Z_{ij}$  and  $W_{ij}$ , and on their comparison. For both the populations  $K = 3$ ,  $k = 2$ ,  $M = 12$  and  $N = 9$ . A unit repeated in two or more clusters represents overlapping. The populations are described in Table 1.

The analysis of the results in Table 1 supports the theoretical developments of the present paper. For both the populations, the factor  $F = N_i [(Z_i - \bar{Y}W_i)^2 + (1/n_i - 1/N_i)D_i^2]$  increases with  $N_i$ , resulting in  $MSE(\bar{y}_{RP}) < MSE(\bar{y}_{RS})$ , as remarked above.

**CONCLUSION**

This paper removes the realistic limitation of known population size in the earlier work of Singh (1988) while considering overlapping clusters. Also comparison of the two strategies here is more direct, whereas in Singh (1988) the support of evidence given by Hansen and Hurwitz (1943) was needed.

**ACKNOWLEDGEMENTS**

This research was partially supported by NSERC Grant A-3111. The comments of the referees and the editor on an earlier version were most helpful in improving the paper. These are gratefully acknowledged.

**REFERENCES**

AGARWAL, D.K., and SINGH, P. (1982). On cluster sampling strategies using ancillary information. *Sankhyā*, B, 44, 184-192.

AMDEKAR, S.J. (1985). An unbiased estimator in overlapping clusters. *Bulletin of the Calcutta Statistical Association*, 15, 231-232.

GIFFARD-JONES, W. (1993). The doctor game. *The Windsor Star*, April 15, 1993.

GOEL, B.B.P.S., and SINGH, D. (1977). On the formation of clusters. *Journal of the Indian Society of Agricultural Statistics*, 29, 53-68.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

SÄRNDAL, C-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, S. (1988). Estimation in overlapping clusters. *Communications in Statistics, Theory and Methods*, 17, 613-621.