MEASUREMENT OF MAIL VOLUMES - AN APPLICATION OF MODEL ASSISTED ESTIMATION

Stephen Woodruff, David W. Hall, Charles Dennis Sissel

Key Words: Ratio Estimation, Regression Model.

Abstract: The estimation of mail volumes (total numbers of pieces) from foreign countries can be difficult under a complex sample design and the simple expansion estimator. This stems from lack of information about daily mail volumes prior to actual sampling and this leads to extreme variations in probabilities of selection. This means that the simple expansion estimator (Horvitz-Thompson) has an unacceptably large variance. Fortunately, mail populations afford other estimation approaches that do work. They are based on physical properties of mail flows and the stochastic structure these impose on samples of mail. This paper outlines sampling designs and estimators that exploit this structure and manifestly reduce mean square error compared to more traditional designs with the expansion estimator.

1. Introduction

The United States Postal Service requires periodic estimates of mail volumes (total pieces/items of mail) for mail entering United States. Since it would be prohibitively expensive to maintain actual counts of pieces for the many thousands of mail subcategories, these counts are estimated through regular sampling of inbound mail flows. These estimates are essential for proper management of mail processing resources and for settlement with foreign postal administrations. Settlement is the process by which the USPS pays these foreign postal administrations (foreign countries) for delivering US origin mail within their borders and in turn gets paid for delivering their mail to US addresses.

Measuring mail volumes for mail entering the US poses some special problems. These problems are at least partially offset by fortuitous properties of mail flows which permit the use of sampling and estimation techniques that cannot be applied to most other survey sampling problems (sampling households, businesses, people). In this paper we describe how to exploit these properties and how to minimize the problems inherent in sampling the mail.

The target population is partitioned into distinct flows (strata) defined by country of origin, transportation mode, arrival city in the US, mail class, reference period (month, quarter, or year), and receptacle type (bag, tray, or tub). Mail entering the US is almost always in some type of receptacle, a bag, a tray, or a large bulk container filled with small packages. These mail flows are the sampling strata. These receptacles are the ultimate sampling units and when selected for sampling, piece counts are recorded by mail class for the entire contents. Our auxiliary variable is weight and this is known for all sample data and total weight is also known for each flow (stratum). This target population poses some interesting sampling problems:

1) A mail receptacle only exists for a few hours.
2) The actual frame or list of receptacles for the study population is rarely completely known and never available for sample design.
3) Only surrogate frame data is available for sample design and this data is not accurate.
4) Due to fickle day-to-day mail volumes and fixed data collection resources, probabilities of selection cannot be controlled or even known in advance of actual data collection.
5) The overwhelming priority for the USPS is speeding mail delivery. Sampling of the mail must not impede delivery.

Historically, standard stratified multi-stage sample designs are used to measure mail volumes. Sampling strata were defined by country of origin (or destination), transportation mode, mail class, month, and office or facility where the mail enters the US (exchange office).

Within these strata, a cluster sample of days was selected and for each sample day, the mail was implicitly stratified by time of arrival. When a dispatch or group of receptacles arrived, a sample of receptacles was selected and the total number of receptacles in the dispatch/group was recorded. By this means the sample data could be expanded to the dispatch/group, then to the day, and finally to the reference period. This Horwitz-Thompson estimator is the standard for such sample designs.

Sampling within a sample day is stratified unpredictably by the groups/dispatches that arrive or depart on the sample day. These last stages of receptacle sampling can and do generate widely differing probabilities of selection. When these probabilities of selection cannot be controlled, there are special risks for Horvitz-Thompson estimation (Basu, 1971).

This paper outlines sampling and estimation strategies that circumvent problems with a multi-stage stratified sample design and the Horvitz-Thompson estimator where probabilities of selection cannot be controlled. These strategies depend on the fact that the number of mail pieces in a mail flow is roughly proportional to the total kilograms of mail in that flow. An analogous statement can be made for numbers of pieces in a receptacle from that flow, or in a sample of receptacles from that flow. Number of pieces (items) divided by their weight is approximately constant for any random selection from a mail flow. This fact justifies the separate ratio estimator and also implies that it is unbiased without regard to sample size.

II. A Model for Mail Flows

The study population of international mail entering the US is called "Letter Post" and consists of letters, cards, flats (magazines, journals), and small packets. A piece of Letter Post cannot exceed 2 Kilograms. It is this limited size for Letter Post items that assures the rough proportionality between pieces and weight for the contents of a mail receptacle. The average mail receptacle (letter trays, flat tub, or bag) is about 8 to 10 Kilograms but can be as light as 1 Kilogram and as heavy as 30 Kilograms.

Letter Post is placed into these receptacles with the single aim of filling them and dispatching them as quickly as possible. Within receptacle type, this process is chaotic with respect to piece weight. The mail pieces in a receptacle can be treated (and historically have been treated) as a simple random sample of mail pieces from the entire mail flow. Within a mail flow (or stratum), it wouldn't be inappropriate to think of the n mail pieces in a given receptacle as the first n pieces in a random sort of the entire flow. Given this operational feature of mail receptacles the following computations help establish the linear regression model ($y = \beta x + \varepsilon$) relating pieces (items) and weight.

Let z_i be the weight in kilograms of the i^{th} mail piece in the mail flow being sampled. Let $\bar{z}$ be the sample mean for the weights of mail pieces in a receptacle from that flow. Then, assuming these are a simple random sample from the entire flow, $E(\bar{z}) = \bar{Z}$, the average piece weight for the entire mail flow. The number of items per kilogram (IPK) for the entire flow is: $1/\bar{Z}$. The quantity to be estimated is this flow IPK times the known flow weight, $1/\bar{Z} \times Z$. Estimating total flow pieces is equivalent to estimating flow IPK, $1/\bar{Z}$.

The IPK for the receptacle is $1/\bar{z}$.

If this receptacle IPK, $1/\bar{z}$, is expanded about $\bar{Z}$ in a Taylor Series:

$$1/\bar{z} \doteq \sum_{i=0}^{L} \frac{(\bar{z}-\bar{Z})^i}{\bar{Z}^{i+1}}(-1)^i \quad \text{and}$$

$$E(1/\bar{z}) \doteq \sum_{i=0}^{L} \frac{E(\bar{z}-\bar{Z})^i}{\bar{Z}^{i+1}}(-1)^i =$$

$$(1/\bar{Z})\left(1+\sum_{i=1}^{L} \frac{E(\bar{z}-\bar{Z})^i}{\bar{Z}^{i}}(-1)^i\right).$$

By Jensen's inequality we know that the relative bias of the sample IPK:

A= $\sum_{i=1}^{L} \frac{E(\bar{z}-\bar{Z})^i}{\bar{Z}^{i}}(-1)^i \geq 0$, and by using historical sample data this term is estimated for small values of L and does indeed get small quickly as n , the number of mail pieces in the receptacle increases.

Table 1 provides estimates of A for L=3. It was also verified that the tail of A for large L converges quickly to zero so that L=3 captures all but a tiny residual portion of the total relative bias.

Table 1. Values of A as a function of n for L=3

n	A
100	.066
200	.044
300	.036
400	.028
500	.023

A single letter tray can contain over 400 pieces of mail and the typical sample size for a flow will be from several to several dozen receptacles. Just as a single receptacle looks like a random selection of mail pieces, the aggregate of mail pieces in a random sample of receptacles looks like a random selection of pieces from the flow and the expectation of $\bar{z}$ is $\bar{Z}$.

The sampling variance of $\bar{z}$ under the design that first selects a random sample of receptacles and then pools all the mail pieces (say m pieces) within these receptacles is approximately the same as the sampling variance of a simple random sample of m mail pieces from the flow because the within receptacle correlation (Intracluster correlation) of piece weights is negligible

This implies that the sample IPK tends to the flow IPK with a bias that rapidly approaches zero as the number of sampled receptacles increases beyond a very modest number (1 for letter trays to 5 for bags). Multiplying the sample IPK by the sample weight:
(sample_pieces)=(population_IPK)x(sample weight) +noise , or ($y = \beta x + \varepsilon$), the model that drives most of the results in this paper.

By the Central Limit Theorem, if piece weights within a mail flow, $\{z_j\}$ are iid with a mean $\bar{Z}$ and variance σ^2, then for a sample of size n > 100 say, $\bar{z} \approx N(\bar{Z}, \sigma^2 / n)$. The term A above, for L=2, or 3 is simply the relative variance of $\bar{z}$. Table 1 provides five estimates of $\sigma^2 / \bar{Z}^2$ and their average is 8.15. When n > 815 , A < .01 and the sample IPK is essentially unbiased. Estimates of the mail flow relative variances, $\sigma^2 / \bar{Z}^2$, are fairly stable when compared across different mail flows. A conservative estimate for $\sigma^2 / \bar{Z}^2$ would be 20 for most mail flows and this estimate could be used with (2.1) below to derive a generalized variance function.

If the sample IPK is approximated for L=3 with:

$$1/\bar{z} \doteq \sum_{i=0}^{3} \frac{(\bar{z} - \bar{Z})^i}{\bar{Z}^{i+1}} (-1)^i$$ then under normality of $\bar{z}$ the variance and bias of the sample IPK are:

$$V\left(1/\bar{z}\right) = \frac{\sigma^2}{n\bar{Z}^4} + \frac{8\sigma^4}{n^2\bar{Z}^6} + O\left(1/n^3\right) \qquad (2.1)$$

and

$$B\left(1/\bar{z}\right) = \frac{\sigma^2}{n\bar{Z}^3}. \qquad (2.2)$$

Then ratio of bias squared to MSE is:

$$\frac{B^2\left(1/\bar{z}\right)}{MSE\left(1/\bar{z}\right)} = \frac{1}{9 + \left(n/R(z)\right)},$$ where

$$R(z) = \frac{\sigma^2}{\bar{Z}^2}$$ and from above, $\hat{R}(z) = 8.15.$

This establishes that bias squared is an inconsequential part of MSE for anything more than very modest sample sizes (in receptacles which each contain many mail pieces). Recall that a single receptacle contains several dozen to several hundred mail pieces and a typical stratum sample is several to several dozen receptacles. n is more than 1000 for most important mail flows, indeed usually much more.

From 2.1, the standard error of a stratum IPK estimate is approximately directly proportional to the stratum IPK and inversely proportional to $\sqrt{n}$. Likewise, from 2.2 the bias of a stratum IPK estimate is directly proportional to the stratum IPK and inversely proportional to n. As n gets large,

$$B\left(1/\bar{z}\right) = o\left(\sqrt{V\left(1/\bar{z}\right)}\right).$$

Another way of looking at mail that yields a similar result posits that a single receptacle is selected randomly from among all the receptacles of a mail flow. Suppose this receptacle weighs x kilograms. The mail in this receptacle can be viewed as a single randomly selected collection or stack of mail from among all the x kilogram stacks of mail that make up the entire flow. Let y_s be the number of pieces in this randomly

selected stack. Because x is a relatively small number, generally around 8 to 10 Kilograms, and Letter Post items are also small, the entire flow can be partitioned into stacks of roughly x kilograms (say N such stacks). If the i^{th} stack has y_i pieces of Letter Post, then $E(y_s) = \frac{1}{N}\sum_{i=1}^{N} y_i$, the population mean. However, $E(y_s)$ is the average number of pieces per x kilograms in the population, or the average number of pieces per kilogram (IPK) in the population times x. This is again the proportionality between pieces and kilograms that is suggested initially above and this relationship is the foundation for unbiased ratio estimation (without regard to sample size in receptacles).

This discussion is based on a partitioning by the auxiliary variable that cannot be assumed for other commonly sampled populations like business establishments, households, or people. If pieces of Letter Post were not restricted to be less then 2 kilograms this argument would fail because the rough partitioning of the flow into piles of about x kilograms (3<x<12) would become a little too approximate for credibility.

Looking at the sample (or receptacle) IPK from a different direction, suppose the regression superpopulation model relating pieces and weight is given. It is shown below that the bias of the sample IPK as an estimate of the population IPK is zero without regard to sample size.

If v and w are two random variables with means V and W respectively, then

$$Cov\left(v/w, w\right) = E(v) - E(v/w)E(w)$$ and rearranging:

$$E(v/w) = \frac{V}{W} - \frac{1}{W}Cov(v/w, w)$$

This states the bias of v/w for estimating V/W is a multiple of $Cov(v/w, w)$. Now assume that $v = \beta w + \varepsilon$, where β is an unknown constant and ε is a random variable when $E(\varepsilon) = 0$. This models the relationship between pieces and weight in mail receptacles. Then conditioning on w (weight) :

$$Cov(v/w, w) = E\left(Cov(v/w, w|w)\right) + Cov\left(E(v/w|w), E(w|w)\right)$$

$$=0+ Cov\left(\beta w/w, w\right) = 0$$ and the bias of the ratio, as an estimate of the ratio of the expected values, is zero. For our purposes, the sample IPK for a mail flow (stratum) is an unbiased estimate of the overall flow IPK.

The way mail is placed in receptacles will tend to assure that the covariance between sample weight and sample IPK is small. This is a consequence of either one of the two things listed below. For mail receptacles both hold and reinforce each other to ensure a small correlation.

1) Mail receptacles within a stratum tend to weigh about the same - what one worker can easily and safely lift.

2) Within detailed strata, mail receptacles are filled without regard to the study variable so that each receptacle looks like a random sample of the mail for that stratum independent of the receptacle's weight.

An unbiased estimate of total pieces for the mail flow is the product of the flow's sample IPK and the total flow weight. The sum of these flow piece estimates is the separate ratio estimator for total pieces and it follows from the above discussion that this estimate is nearly unbiased for modest stratum sample sizes (of receptacles). This is the basis for reducing variance by using extremely detailed stratification (many strata) together with the separate ratio estimator. Large numbers of strata may only be possible with a very small sample size in each stratum(flow).

The separate ratio estimator is also the Best Linear Unbiased Estimator (BLUE) under the established superpopulation model with a few additional second moment assumptions on the residuals.

III. An Efficient Sample Design

Multi-stage cluster sample designs are sometimes necessary due to sample size constraints. If a study population is divided into groups and there are more of these groups than the total affordable sample size of ultimate sample units, then it may be necessary to select a random sample of groups

(clusters) and then sub-sample each of the selected clusters. This is a single stage cluster sample.

If the ultimate sample size is larger than the number of groups and particularly if the population units within each group are relatively homogeneous with respect to the study variable(s) then a stratified design will be more efficient. The groups become strata and a small sample is selected from each stratum. The stratified Horvitz-Thompson estimator (HT) will have smaller variance under this design than the HT under the clustered design.

Table 2 demonstrates the effect on sampling variability of moving from a deeply stratified design with a sample of size one in each of 128 strata through a series of two stage cluster designs. Each of these cluster designs ultimately selects the same number of sample receptacles, 128. As the sample size of cluster decreases, the number of sample receptacles per cluster increases to always maintain a final sample at 128 receptacles.

For example, the third row in Table 2 (64 Com) has results for the design that selects a first stage sample of 64 clusters (SRS WOR) from the 128 that partition the population. Then 2 receptacles are selected (SRS WOR) from each of these 64 sampled clusters and the population total is estimated with the combined ratio estimator as described in detail in the Appendix. The Variance Ratio for row 64 Com is .637, the variance under 128 Sep divided by the variance under 64 Com. Row 32 Com describes results for the combined ratio estimator applied to 32 sampled clusters with 4 receptacles per cluster and so on.

Table 2.

Number of Clusters Sampled	Sample size per cluster	Variance Ratio *	Relative Error **
128 Sep	1	1	15.0%
128 Com	1	.96	15.4%
64 "	2	.637	18.8%
32 "	4	.382	24.3%
16 "	8	.212	32.7%
8 "	16	.112	44.9%
4 "	32	.058	62.6%
2 "	64	.029	88%
1 "	128	.015	124%

* Variance ratio is the variance of the separate ratio estimator with 128 strata divided by the variance of the combined ratio estimator for the row's cluster design.

** Relative Error is the relative error of the estimator for the row's design. See Appendix for the stochastic structure and the formulae used to derive Table 2.

These variance computations are made from historical sample data for inbound Letter Post. For the stratified design with a single sample receptacle per stratum, the separate ratio estimator (Sep) of total mail pieces was used. For the cluster designs, the combined ratio estimator was used (according to historical practice in the USPS).

Combining the discussion in this section with that in section II, the best sampling and estimation strategy is extremely detailed stratification and the separate ratio estimator. In the next section it is demonstrated that the variance component of MSE accounts for the lion's share of the MSE (95% to over 99%) even for a minimum stratum sample size. The usual way to reduce variance, by increasing sample size, would virtually eliminate the bias completely. There are other administrative and operational advantages in this strategy that may be almost as important as the gains in precision.

IV Simulation Study with Actual Sample Data and Data Sets Generated From This Data

The statements above about bias in the separate ratio estimator may need further justification since they are based on approximations. The primary purpose of this simulation study is verification of the claim that the bias of the separate ratio estimator for mail volumes is negligible for small stratum sample sizes (in receptacles). A secondary purpose is to show that the separate ratio estimator is more robust than the combined ratio estimator against some common difficulties of sampling mail. This study uses actual populations of mail receptacles that were sampled coming into the US during 2003.

All inbound mail receptacles for 82 mail flows (or strata) of Letter Post mail (letters, cards, magazines, and small packages) that were sampled in our inbound data collection system were used for the test population. These 82 strata are each defined by Country of origin, Transportation mode (air, surface, SAL), office of entry into the US, and receptacle type (letter tray, flat tray, bag). These stratum sizes varied from

one receptacle to several hundred (and for generated populations, several thousand).

In order to test for bias in the separate ratio estimator, the sample size in each stratum was a single receptacle. This should be the worst case (maximum bias) for the separate ratio estimator.

Three estimators for total pieces in this test population were compared, the Horvitz-Thompson estimator (HT), the combined ratio estimator (HTC) {=[HT(pieces)/HT(Kilograms)]x(known population Kilograms)}, and the separate ratio estimator (SRE).

500 stratified simple random samples were selected and for each of these samples, HT, HTC, and SRE were computed. These 500 replicates of the three estimators and the known population piece total were used to estimate the Root Mean Square Error (RMSE) and percent of MSE that comes from squared bias (% Bias). The results of some of these simulations are summarized in the Table 2 directly below.

Table 3.

Population	Expected Error	HT	HTC	SRE
A	RMSE	347	328	325
	% Bias	4.1	3.0	5.6
B	RMSE	2414	1808	1145
	% Bias	0	.8	2.0
C	RMSE	1784	1402	1070
	% Bias	.01	.01	0.1
D	RMSE	3093	1898	1145
	% Bias	0.2	.09	.02

Population A is the population of all 2003 sample receptacles in the 82 strata. Note that for Population A, all three estimators are similar for RMSE and % Bias. Since we know that HT is unbiased, 4.1 reflects the order of magnitude for a "statistical zero" in this simulation. The % Bias for HTC and SRE are similar to the %Bias of HT.

Population A may be too well behaved to reflect the large volume differences between strata (sample sizes in different strata tend to be far more similar than their respective strata population sizes) and occasionally bulk receptacles that are found in some sampling strata. When a bulk receptacle is encountered in a bag stratum, the bulk receptacle is sub-sampled and one or two receptacles inside are selected and their data expanded to estimate for the totality of mail in the bulk receptacle. Populations B, C, and D are different simulated populations derived from A by generating artificial receptacles for some of the 82 strata. These additional mail receptacles were generated from distributions that resemble the empirical sampling distribution except that some large (bulk) receptacle were generated that had IPKs similar to other receptacles in the stratum. For these populations, the estimated bias of SRE was also negligible but its MSE was much smaller than the two competitors.

V. Estimation of Other Mail Characteristics.

This discussion has focused on the measurement of total items in a mail flow by estimation of IPK (items per kilogram) for the flow. If w is any other quantifiable characteristic of a mail piece, then a very similar approach can be used to estimate the average (or total) of that characteristic for the entire flow.

The sample mean of w over all mail pieces sampled for the flow, $\overline{w}$, is normally distributed and an unbiased estimate of $\overline{W}$, the population mean for the flow. $\overline{w}$ estimates the average per piece value of w in the flow. Thus an estimate of the population total of w in the flow is:

$$\hat{T}_w = \overline{w} \times (sample\ IPK) \times Z = \overline{w}\left(1/\overline{z}\right)Z,$$

where Z is the known total flow kilograms and $\overline{z}$ is, as in section 2.2, the sample average piece weight. $\hat{T}_w$ is the sample estimate of w per kilogram of mail times the total flow kilograms. For example, if w is one for mail pieces going to New York and zero otherwise then $\hat{T}_w$ is the estimate of total number of mail pieces going to New York in the mail flow being considered. w could also be something related to the postage paid on a piece of mail.

This estimator can be analyzed (mean, variance, etc) by applying normal theory to its Taylor approximation to find its analogs to (2.1) and (2.2). By this means, the estimation of flow total for any characteristic associated with a mail piece is routine.
For example let:

$$\begin{pmatrix} \overline{z} \\ \overline{w} \end{pmatrix} \propto N\left(\begin{pmatrix} \overline{Z} \\ \overline{W} \end{pmatrix}, \frac{1}{n}\begin{pmatrix} \sigma_z^2 & \sigma_{zw} \\ \sigma_{zw} & \sigma_w^2 \end{pmatrix}\right)$$

If a first order Taylor approximation to $\frac{\overline{w}}{\overline{z}}$ is used then:

$$\frac{\overline{w}}{\overline{z}} \propto N\left(\left(\frac{\overline{W}}{\overline{Z}}\right), \frac{\overline{W}^2}{\overline{Z}^2}\frac{1}{n}\left(\frac{\sigma_w^2}{\overline{W}^2}+\frac{\sigma_z^2}{\overline{Z}^2}-2\frac{\sigma_{zw}}{\overline{Z}\overline{W}}\right)\right)$$

where n is the number of mail pieces in the flow sample. The variance is a function of R(z), R(w), and R(zw) somewhat similar to (2.1). Then the variance of $\hat{T}_w$ is:

$$V(\hat{T}_w) = V\left(\frac{\overline{w}}{\overline{z}}Z\right) =$$

$$Z^2\left(\frac{\overline{W}^2}{\overline{Z}^2}\frac{1}{n}\left(\frac{\sigma_w^2}{\overline{W}^2}+\frac{\sigma_z^2}{\overline{Z}^2}-2\frac{\sigma_{zw}}{\overline{Z}\overline{W}}\right)\right).$$

VI. Conclusions.

Sampling for mail volumes coming into the US poses unique problems but these problems can be overcome by exploiting the stochastic structure imposed on mail flows by the way mail is processed. A traditional sample design and its designed based estimator can play to the weaknesses of postal data systems if some care is not taken with the sample design. A model assisted estimation methodology exploits inherent properties of postal data and improves the precision of the volume estimates.

One problem in sampling mail is the difficulty of constructing a good sample frame in time for sample selection which is usually several months before the actual sample receptacles are chosen and their data recorded. There is no possibility of knowing on any giving future day how much mail will enter the USPS processing and distribution system until after it has come and gone.

Complex sample designs must depend on surrogate frame data that is often much different from the actual frame data that is recorded as part of data collection. This results in probabilities of selection that cannot be controlled and an extremely wide range of expansion factors. Such designs can yield estimators with very large variances - a problem which has been dealt with by routine outlier adjustment.

The solution proposed here is to sample the mail with a simple single stage finely stratified sample design. The estimator of choice is the separate ratio estimator. This estimator is unbiased even for tiny stratum sample sizes, robust against the frame difficulties inherent in mail sampling, and administratively simpler. It has smaller MSE than either the Horvitz-Thompson estimator or the combined ratio estimator.

Historically, this strategy may have been rejected due to fears of bias when using a separate ratio estimator and small stratum sample sizes. The stochastic structure of mail flows as described in detail in section II implies that such fears are unfounded. The real problem with the combined ratio estimator and complex designs was excessive variance (as described above).

There are other solutions to the problem of extreme variation in the sample expansion factors but the one proposed here is quite simple and has administrative advantages that may be as important as MSE reduction.

This type of sampling problem must have occurred in sampling biological populations but I'm not familiar with the literature or terminology that may be available on this subject. My apologies to others who have done similar work but who have not been referenced here.

Appendix

The computations in Table 2 are based on the variance formula and stochastic structure given in this appendix.

Suppose a population is divided into M clusters. Suppose a sample, s, of size n of these M clusters is selected as the first stage of a cluster sample. Let s be a simple random sample without replacement (SRS WOR) and N_i be the population size of the i^{th} cluster. When n=M this becomes a stratified sample. Let Y_{ij} be the study variable for the j^{th} member of the i^{th} cluster (stratum when n=M) and X_{ij} is the auxiliary variable for the j^{th} member of the i^{th} stratum. $Y_i = \sum_{j=1}^{N_i} Y_{ij}$ and $X_i = \sum_{j=1}^{N_i} X_{ij}$ and X_i

is known for all M clusters. Let s_i be a simple random sample without replacement from the i^{th} cluster of size $n_i < N_i$.

The separate ratio estimator for the population total is:

$$\hat{T}_S = \frac{M}{n}\sum_{i\varepsilon s}\frac{N_i \bar{y}_{s_i}}{N_i \bar{x}_{s_i}} X_i ,$$ where

$\bar{y}_{s_i} = \frac{1}{n_i}\sum_{j\varepsilon s_i} y_{ij}$, $\bar{x}_{s_i} = \frac{1}{n_i}\sum_{j\varepsilon s_i} x_{ij}$, and lower case is used for sample values of the X and Y variates. Let $y_{s_i} = n_i \bar{y}_{s_i}$ and similar for x_{s_i}.

The combined ratio estimator for the population total is:

$$\hat{T}_C = \frac{\sum_{i\varepsilon s}\frac{M}{n}\frac{N_i}{n_i} y_{s_i}}{\sum_{i\varepsilon s}^{M}\frac{M}{n}\frac{N_i}{n_i} x_{s_i}} \sum_{i=1}^{M} X_i$$

Let $S_{R_i} = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \frac{Y_i}{X_i} X_{ij})^2$

$$S_{O_i} = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}([Y_{ij} - \frac{Y}{X} X_{ij}] - \overline{O_i})^2$$

where $\overline{O_i} = \frac{1}{N_i}\sum_{j=1}^{N_i}\left[Y_{ij} - \frac{Y}{X} X_{ij} \right]$, and $Y = \sum_{i=1}^{M} Y_i$, and $X = \sum_{i=1}^{M} X_i$

Let $S^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(Y_i - \frac{Y}{X} X_i \right)^2$ The variances of the separate ratio estimator for n=M (stratified sampling case) is:

$$V\left(\hat{T}_S\right) = \sum_{i=1}^{M}\frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) S_{R_i} .$$

The variance of the combined ratio estimator for both M=n and M>n is:

$$V\left(\hat{T}_C\right) = \frac{M}{n}\sum_{i=1}^{M}\frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right) S_{O_i}^2 + \frac{M^2}{n}\left(1 - \frac{n}{M}\right) S^2 .$$

References:

Basu, D. (1971) "Foundations of Statistical Inference", Edited by Godambe V. P. and Sprott D. A., Toronto, Holt, Rinehart, and Winston. pp203 - 24 2