REMAINDER LINEAR SYSTEMATIC SAMPLING WITH MULTIPLE RANDOM STARTS

By

SAYED A. MOSTAFA ABDELMEGEED

Bachelor of Science in Statistics

Cairo University

Cairo, Egypt

2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE

May 2014

REMAINDER LINEAR SYSTEMATIC SAMPLING
WITH MULTIPLE RANDOM STARTS

Thesis Approved:

Dr. Ibrahim A. Ahmad
Thesis Advisor

Dr. Carla Goad
Commitee Member

Dr. Ye Liang
Commitee Member

ACKNOWLEDGMENTS

I would like to thank professor Ibrahim A. Ahmad for his guidance and advice through the preparation of this work. I also would like to thank my committee members, professors; C. Goad and Y. Liang for their valuable comments.

I am also thankful to professors; Laila O. El-Zeini and Ramdan Hamd for their encouragement in the early stages of this work.

Last but not the least, I wish to thank my parents, brothers, sisters, and my wife for their support during the preparation of this work.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: SAYED A. MOSTAFA ABDELMEGEED

Date of Degree: MAY, 2014

Title of Study: REMAINDER LINEAR SYSTEMATIC SAMPLING WITH MULTIPLE RANDOM STARTS

Major Field: STATISTICS

Abstract

"Systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling" (Levy (2008) p.83). This fact is due to the simplicity and the operational convenience of this technique. However, this technique has two main statistical problems. First, if the sampling interval, $k = N/n$, is not an integer, the actual sample size will not be fixed and the sample mean, $\bar{y}$, will not be unbiased estimator for $\bar{Y}$, the population mean. Second, regardless of the sampling interval, the sampling variance of the estimator $\bar{y}$ cannot be consistently estimated on the basis of a single systematic sample. In this study, we introduce a new generalized systematic sampling design that can handle these two issues simultaneously. The proposed design is a generalization of the remainder linear systematic sampling design of Chang and Huang (2000), which handles only the problem of non-integer sampling intervals. Unbiased estimators for both $\bar{Y}$ and the sampling variance are derived under the proposed design. The performance of the proposed design is evaluated in comparison to five sampling procedures under different supperpopulation models. Specifically, simple random sampling, remainder linear systematic sampling, circular systematic sampling, new partially systematic sampling and mixed random systematic sampling. It is found that our proposed design performs well compared to the other designs in most cases.

Contents

Glossary

BSS Balanced Systematic Sampling - Sethi (1965)

BSSM Balanced Systematic Sampling with Multiple Random Starts - Sampath (2010)

CSS Circular Systematic Sampling - Lahiri (1951)

LSS Linear Systematic Sampling - Madow and Madow (1944)

MRS Multiple Random Starts

MRSS Mixed Random Systematic Sampling - Huang (2004)

MSS Modified Systematic Sampling - Singh *et al.* (1968)

MSSM Modified Systematic Sampling with Multiple Random Starts - Sampath (2010)

MSSS Multi-Start Systematic Sampling - Gautschi (1957)

NPSS New Partially Systematic Sampling - Leu and Tsui (1996)

NSS New Systematic Sampling - Singh and Singh (1977)

PSS Partially Systematic Sampling - Zinger (1980)

RLSS Remainder Linear Systematic Sampling - Chang and Huang (2000)

RLSSM Remainder Linear Systematic Sampling with Multiple Random Starts

SRS Simple Random Sampling

STRS Stratified Random Sampling

1

Introduction

Sampling design in which only the first unit is randomly selected, the rest being automatically selected according to a predetermined pattern is known as systematic sampling. Systematic sampling is one of the most prevalent sampling techniques. Its popularity is mainly due to its practicability. Compared to simple random sampling, it is easier to draw a sample especially when the drawing is done in the field. In addition, systematic sampling can provide more precise estimators than simple random sampling when explicit or implicit stratification is present in the sampling frame (Cochran 1977). This is due to the fact that systematic sampling stratifies the population into n strata each of size k units and selects one unit from each stratum. So, systematic sampling is expected to be about as precise as the corresponding stratified random sampling with one unit per stratum. It is also efficient in sampling some natural population like forest areas for estimating the volume of timber (Zinger 1964). Thus, many research centers use the systematic sampling design in their surveys. For example, the Food and Agriculture Organization (FAO) of the United Nations utilizes systematic sampling in conducting its Global Forest Resources Assessment Survey (GFRAS 2010).

In its simplest form, called linear systematic sampling (LSS), the systematic design can be described as follows. In order to choose a systematic sample of size n from a population of size N, the population is first divided into n groups each of size k units, where $k = N/n$ is called the sampling interval. A random start r is chosen from the first k-units group. All units corresponding to r in the remaining groups are then chosen in the sample. The obtained sample will hence comprise the units with indices;

$$\{r, r+k, r+2k, \ldots, r+(n-1)k\}.$$

For instance, if $N = 40$ and $n = 10$, then using $k = N/n = 4$ a random start r is first chosen from the first 4 units in the population. Let $r = 3$, the second observation in the sample will be the unit with index $r+k = 7$, the third will be the unit with index 11 and so on. The sampled units in this example will be those units with indices corresponding to;

$$\{3, 7, 11, 15, 19, 23, 27, 31, 35, 39\}.$$

The systematic design is an equal probability of selection method. However, when the population size N is not a multiple of the sample size n, the sample size will not be fixed and the sample mean becomes a biased estimator of the population mean (Kish 1965). In addition, the properties of the estimators from systematic samples depend on the order of the units in the frame, and can be less efficient under some arrangements. For instance, the existence of a linear or parabolic trend could produce less precise estimators. Also, if the population consists of a periodic trend like a sine curve, and the sampling interval, k, is equal to the period of the curve or an integral multiple of the period, the efficiency of a systematic sample of size n will be very close to that of only one observation taken randomly from the population. If k is an odd multiple of the half-period, the systematic sample mean will coincide with the true population mean (Cochran 1977).

Linear systematic sampling is first introduced by Madow and Madow (1944). It can be viewed as a cluster sampling where only one cluster is chosen randomly from k clusters each of size n units. Therefore, a single systematic sample alone cannot be used to estimate the standard error of the sample mean and other sample statistics (Chaudhuri and Stenger 2005). One common practice in applied surveys is to regard LSS as a simple random sample. However, such practice typically provides highly biased estimators of the sampling variance (Wolter 1984). If one is reasonably aware of the underlying model of the sampled finite population or, alternatively, the assumed infinite superpopulation, more appropriate estimators for the systematic sampling variance can be derived. Wolter (1984) reviewed and compared eight biased estimators for the sampling variance. Specific guidelines were then introduced based on the comparison of the mean square errors of the eight estimators under various models. Two specific estimators have been signaled as good general-purpose estimators when little is known about the population. These two estimators, denoted by v_2 and v_3, take the following forms:

$$v_2 = \frac{(1-f)}{2n(n-1)} \sum_{j=2}^{n} (y_j - y_{j-1})^2 \tag{1.1}$$

$$v_3 = \frac{(1-f)}{n^2} \sum_{j=2}^{n/2} (y_{2j} - y_{2j-1})^2 \tag{1.2}$$

Both estimators treat the systematic sample as a special case of the stratified sample, where each stratum has $2k$ units and two units are chosen from each stratum using simple random sampling without replacement (SRSWOR). While v_3 is based on non-overlapping differences (i.e. assumes non-overlapping strata), the estimator v_2 is based on overlapping differences and aims at increasing the degrees of freedom (Wolter 1984).

Several modifications to the LSS are introduced to tackle its main two statistical issues, namely, the unfixed sample size in case of non-integer sampling intervals and the difficulty of estimating the sampling variance. These modifications are reviewed below. There is still further room for modifying the LSS to deal with these two shortcomings. The proposed study introduces such approach. In this study, the focus will be on remainder linear systematic sampling (RLSS) of Chang and Huang (2000) where the idea of multiple random starts will be incorporated into this design in an attempt to get an

unbiased estimator for the variance of the sample mean corresponding to this design. Hence, the new developed design will be called remainder linear systematic sampling with multiple random starts (RLSSM).

Briefly, this study aims to achieve the following three main objectives:

I. Obtaining an unbiased estimator for the population mean under the proposed design.

II. Estimating the variance of the sample mean unbiasedly under the proposed design.

III. Investigating the performance of the proposed design relative to some other systematic sampling designs, namely, simple random sampling (SRS), circular systematic sampling (CSS), RLSS, new partially systematic sampling (NPSS) and mixed random systematic sampling (MRSS) through a comparative study. The comparisons will be carried out numerically in cases where the performance cannot be mathematically studied. Since the relative efficiencies depend on the population structure, the performance comparisons will be done under different superpopulation models.

The remaining chapters of the current work are organized as follows: Chapter 2 will be devoted to reviewing background and literature. The proposed design, the estimators and their statistical properties (unbiasedness) are presented in Chapter 3. Chapter 4 includes performance comparisons of the proposed design with some other sampling designs under different superpopulation models. Conclusions and suggestions for future research are presented in Chapter 5.

2

Literature Review

This chapter introduces the systematic sampling design relative to the other common designs, namely, stratified sampling and SRS (Section 2.1). It also discusses some of the available approaches for dealing with the statistical problems of the systematic sampling design (Sections 2.2 - 2.5).

2.1 Systematic Sampling Efficiency Relative to Stratified Sampling and Simple Random Sampling

Cochran (1977) compared the performance of the systematic sampling design with both of stratified and simple random sampling through rewriting the variance of the systematic sample mean in different ways as follows;

$$Var(\bar{y}_{sys}) = (\frac{N-1}{N})S^2 - \frac{k(n-1)}{N}S^2_{wsy} \tag{2.1}$$

$$Var(\bar{y}_{sys}) = \frac{S^2}{n}(\frac{N-1}{N})[1+(n-1)\rho_w] \tag{2.2}$$

$$Var(\bar{y}_{sys}) = \frac{S^2_{wst}}{n}(\frac{N-n}{N})[1+(n-1)\rho_{wst}] \tag{2.3}$$

$$Var(\bar{y}_{srs}) = (\frac{N-n}{N})\frac{S^2}{n} \tag{2.4}$$

and

$$Var(\bar{y}_{st}) = (\frac{N-n}{N})\frac{S^2_{wst}}{n} \tag{2.5}$$

where, S^2_{wsy} is the variance within the systematic samples, ρ_w is the correlation coefficient between pairs of units that are in the same systematic sample, S^2_{wst} is the variance among units that lie in the same stratum, ρ_{wst} is the correlation between the deviations from the stratum means of pairs of items that are in the same systematic sample and $Var(\bar{y}_{st})$ is the variance of the sample mean of stratified random sample with one unit per stratum.

By comparing formulas (2.1) and (2.4), the systematic sample mean will be more efficient than the SRS mean if $S^2_{wsy} > S^2$ i.e. the units within the same systematic sample are heterogeneous. Formula (2.2) shows that the existence of positive correlation between units in the same systematic sample ($\rho_w > 0$) inflates the sampling variance. Hence, systematic sample is less efficient than SRS if there is a positive correlation between units in the same systematic sample. Formula (2.3) relates the variance of systematic sample mean with that of stratified random sample, given in (2.5), through ρ_{wst}. If $\rho_{wst} = 0$, systematic sampling has the same efficiency of stratified sampling with one unit per stratum. If $\rho_{wst} < 0$, systematic sampling is more efficient than stratified sampling with one unit per stratum. Otherwise, stratified sampling is more efficient than systematic sampling.

The previous investigation shows that systematic sampling is sometimes more efficient than both of SRS and stratified sampling and sometimes it is less. In addition, the practicability of systematic sampling makes it more preferable design in many situations even with its other statistical shortcomings.

In the next three sections, different approaches, presented in the literature, for handling the statistical limitations of the systematic sampling design are introduced.

2.2 Procedures for Dealing with Variable Systematic Sample Size

When the population size is not a multiple of the sample size (i.e. $N \neq nk$) the LSS results in a variable sample size and the sample mean becomes biased as an estimator for the population mean. To overcome this drawback, many modifications on the LSS were proposed.

2.2.1 Circular Systematic Sampling

Lahiri (1951) suggested a sampling design where the units of the population are considered to be arranged around a circle. In such case, instead of dividing the population into n groups and selecting a random number from the first group, as in LSS, a random number r is selected between 1 to N. Every k^{th} unit is then chosen in a cyclic manner to be in the sample, where $k = [N/n]$, the integer part of N/n, is the sampling interval. This design is based on the convention that for any $i = 1, 2, \ldots, N$, the unit with index $i + N$ stands for the unit with index i, and hence this design is known as circular systematic sampling (CSS). The units in the sample are those with indices;

$$\begin{cases} \text{r+jk} & \text{if} \quad 1 \leq r + jk \leq N \\ \text{r+jk-N} & \text{if} \quad r + jk > N \end{cases} \quad ; \; j = 0, 1, \ldots, (n-1)$$

Under this design an unbiased estimator for the population mean can be obtained as follows:

$$\bar{y}_{CSS} = \frac{1}{n} \sum_{j=0}^{n-1} y_{r+jk}$$

Sudakar (1978) pointed out that to achieve the required sample size n in CSS, k should be chosen as the largest integer smaller than or equal to N/n.

Sengupta and Chattopadhyay (1987) mentioned that a necessary and sufficient condition to make the CSS of size n, drawn from a population of N units with sampling interval k, contain all distinct units is that $N/(N,k) \geq n$ or equivalently, $[N,k]/k \leq n$, where $(N,k)$and $[N,k]$denote, respectively, the greatest common divisor (g.c.d) and least common multiple (l.c.m) of N and k.

Like the usual LSS, the CSS design does not provide an unbiased estimator for the sampling variance of the sample mean.

2.2.2 Remainder Linear Systematic Sampling

Chang and Huang (2000) proposed another sampling procedure that can be used when $N = nk + r; 0 < r < n$. This procedure is based on the fact that the population size can be written as $N = (n-r)k + r(k+1)$, which means that the population can be divided into two strata, the first consists of the front $(n-r)k$ units, and the second stratum contains the remaining $r(k+1)$ units. A linear systematic sample of size $(n-r)$ units is selected from the first stratum with k as the sampling interval, and another linear systematic sample of size r units is selected from the second stratum with $(k+1)$ as the sampling interval. Combining the two samples together, we get a sample of size n as desired. This design is called remainder linear systematic sampling (RLSS). It is reduced to LSS if the remainder is zero, $r = 0$. Under the RLSS, an unbiased estimator for the population mean can be obtained as follows:

$$\bar{y}_{RLSS} = \frac{1}{N}[(n-r)k\bar{y}_1 + r(k+1)\bar{y}_2]$$

where, $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the first and second stratum, respectively. However, under this design there is no unbiased estimator for the sampling variance.

2.3 Procedures for Estimating Systematic Sampling Variance

As it is mentioned above, a single systematic sample cannot be used solely to estimate the variance of the sample mean, or any other sample statistic of interest, unbiasedly. This issue can be tackled using one of several approaches presented in literature. The following is a review of such approaches.

2.3.1 Model Based Estimators

This approach is based on assigning a model that best characterizes the nature of the values of the variable of interest when they are arranged in certain order. For example, several model-based estimators are given in Cochran (1977, p.223). Each of these estimators is approximately unbiased under a specific underlying model but can be highly

biased under the other models. Therefore, the researcher should be careful while choosing the estimator to be used. The golden rule would be to avoid using LSS if the underlying model of the sampled population is unknown (Wolter (2007)).

Montanari and Bartolucci (1998) proposed a model-based estimator of the variance of the systematic sample mean that is based on a sum of two components. The first component takes into account the trend in the frame of the sampled finite population while the second takes into account the stochastic nature of a general superpopulation model. Given that y_{ij} is the value of the study variable Y of the j^{th} unit in the i^{th} systematic sample, they assumed that y_{ij} is a realization of the random variable Y under the following superpopulation model:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

$$E_M(y_{ij}) = \mu_{ij}, \quad V_M(y_{ij}) = \sigma^2_{ij}, \quad Cov_M(y_{ij}, y_{i'j'}) = 0 \quad \forall i \neq i' \text{ or } \quad j \neq j'$$

They showed that $E_M[Var(\bar{y}_{sys})]$ can be divided into two parts where the first part is due to the systematic component and the other is due to the random component of the assumed model. Based on this idea, they introduced their model based estimator. Their estimator was shown to outperform both the overlapping difference estimator v_2 (equation 1.1) and the simple random sample estimator under several superpopulation models.

Wolter (2007, sec. 8.2.2) proposed a general methodology for constructing a model based estimator for $Var(\bar{y})$. In this methodology, the model dependence is explicitly recognized. The proposed general estimator of the variance is defined as a conditional expectation of $Var(\bar{y})$ given the data y_i from the observed sample;

$$v_i = E[Var(\bar{y})|y_i]$$

where, E denotes the expectation over the assumed specific model.

In this context, Wolter (2007) notes that the practicing statistician must make a professional judgment about the form of the model, as it is never known exactly, and then derive v_i under the selected form. Hence, the variance estimator will be subject to errors of estimation as well as to errors of model specification. Therefore, the applicability of the model based approach is viewed, in practice, as being hampered by lack of robustness.

This lack of robustness can be partly handled by the use of a nonparametric model specification. This class of models, compared to parametric models, makes much less restrictive assumptions on the shape of the relationship between variables which significantly reduces the risk of model misspecification.

Opsomer *et al.* (2012) proposed a new estimator of the model based expectation of

the design variance under a nonparametric model for the population. They derived their estimator under the following superpopulation model:

$$Y_j = m(x_j) + v(x_j)^{1/2} e_j; 1 \leq j \leq N \tag{2.6}$$

where $m(.)$ and $v(.)$ are continuous and bounded functions, x is a univariate auxiliary variable and the errors e_j are independent random variables with mean 0 and variance 1. The design variance of $\bar{y}$ is then written as follows:

$$Var_d(\bar{y}) = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_r - \bar{Y})^2 = \frac{1}{kn^2} Y^T D Y$$

where, $Y = (Y_1, \ldots, Y_N)^T$ and $D = E^T H E$ with $E = 1_n^T \otimes I_k$, $H = I_k - \frac{1}{k} 1_k 1_k^T$, $\otimes$ denoting the Kronecker product[1] and 1_k a vector of ones of length k. The model anticipated variance of $\bar{y}$ under the assumed model, in (2.6), is then defined as follows:

$$E[Var_d(\bar{y})] = \frac{1}{kn^2} m^T D m + \frac{1}{kn^2} tr(D\Sigma)$$

where, $m = [m(x_1), \ldots, m(x_N)]^T$ and $\Sigma = diag[var(x_1), \ldots, var(x_N)]$. This anticipated variance is then estimated using the local polynomial regression.

2.3.2 Modifying the Linear Systematic Sampling Design

Instead of depending on a model based estimator to estimate the systematic sampling variance, many authors suggested to modify the design itself in a way that enables the derivation of unbiased estimators for the sampling variance. Some of these modifications are reviewed below.

a. Mixed Random Systematic Sampling Designs

In this approach, a systematic sample is first chosen from the population and then supplemented with an additional simple random sample without replacement or with another systematic sample from the remainder of the population. The two samples are then used to provide an unbiased estimator for the variance of the estimator of the population mean. This approach, with some variations, is adopted by many authors.

Zinger (1980) introduced the partially systematic sampling (PSS) design in an attempt to provide an unbiased estimator for the sampling variance. This design can be described as follows:

To select a sample of size n from a population of size $N = mk$, where $m < n$, a linear systematic sample of size m units is first chosen. A simple random sample (SRS) of size

[1] Kronecker product $\otimes$: Let $A \in R^{(mn)} and B \in R^{(pq)}$. Then the Kronecker product of A and B is defined as the matrix; $A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in R^{(mp \times nq)}$ (Hadi 1996).

$(n-m)$ units is then selected from the remaining $(N-m)$ units. The final sample will be the union of the two samples. To estimate the population mean $\bar{Y}$, a weighted sample mean can be used as follows; $\bar{y}_w = \alpha\bar{y}_{sys} + \beta\bar{y}_{srs}$, with $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. This procedure appears to provide an unbiased estimator for the sampling variance through the sample sum of squares.

Singh and Singh (1977) proposed the new systematic sampling (NSS) design where choosing a sample of size n involves two steps. First, a sample of u consecutive units is selected by choosing a random number t between 1 and N. A circular systematic sample of size $(n-u)$ units is then selected using sampling interval $k = N/n$. The sampled units are those with indices;

$$\{t+i; i = 0,1,2,\ldots,u-1\} \;\; \& \;\; \{t+u-1+jk; j = 1,2,\ldots,n-u\}.$$

A sufficient condition to guarantee that the sampled units will be distinct is $u+(n-u)k \leq N$. They also mentioned that another sufficient condition which should be added to make all the pairs of units have non-zero inclusion probabilities, and hence the sampling variance can be estimated unbiasedly, is $u \geq k$ and $u+(n-u)k \geq \frac{1}{2}N+1$.

Leu and Tsui (1996) modified the NSS of Singh and Singh (1977) by suggesting new partially systematic sampling (NPSS) design. The new design modifies the NSS by choosing a random sample of size a from the indices;

$$\{t, t+1, \ldots, t+u-1\}$$

where, $u = N-(n-a)k$ and $a = 2$ if $N = nk$; otherwise, $k = [N/(n-1)]$ and $2 \leq a \leq [N/2]+1$. A systematic sample of size $(n-a)$ is then chosen, in a circular manner, which includes the units with indices;

$$\{t+u-1+jk; j = 1,2,\ldots,(n-a)\}.$$

Combining the two samples together, a sample of size n can be formed from which an unbiased estimator for the sampling variance can be obtained provided that $a \geq 2$ and $u \geq k$, through utilizing the second order inclusion probabilities.

Huang (2004) proposed a mixed random systematic sampling (MRSS) design from which both the population mean and the sampling variance can be estimated unbiasedly. Considering the population as if arranged in a circular manner, the proposed design involves two steps. In the first step an index t is selected randomly between 1 and N. Then the population is divided into two subpopulations; the first consists of $(n-r)k$ units with indices $\{t, t+1, \ldots, t+(n-r)k-1\}$ and the second subpopulation contains $r(k+1)$ units with the remaining indices. In the second step, a simple random sample of size $(n-r)$ is drawn from the first subpopulation and a sample of r units, with indices $\{t+(n-r)k-1+j(k+1); j = 1,2,\ldots,r\}$ is selected systematically from the second one. The final sample is then the union of the two samples. If $N = nk$, MRSS will be equivalent to SRS. Under the proposed procedure, the inclusion probabilities were

derived and the Horvitz-Thompson (HT) estimator was used to estimate the population mean.[2] Also, an unbiased estimator for the variance of the HT estimator was presented by utilizing the second order inclusion probabilities.

It is noteworthy that the last three methods, namely, NSS, NPSS and MRSS can also be used when the population size is not a multiple of the sample size ($N \neq nk$) as their introducers, Singh and Singh (1977), Leu and Tsui (1996) and Huang (2004), respectively, mentioned.

b. Multi-Start Systematic Sampling

Gautschi(1957) proposed LSS with multiple random starts aiming at providing an unbiased estimator for the systematic sampling variance. According to this approach, in order to select a systematic sample of size n, one chooses t independent systematic subsamples. For $N = nk$ and n/t is integer, we first choose t random numbers from the front tk units (the first group), say $\{r_1, \ldots, r_t\}$. Then for each chosen random start, a systematic sample is selected by choosing the corresponding tk^{th} units. Finally, the sample will contain units with the following indices;

$$\{r_1, r_1 + tk, \ldots, r_1 + (\tfrac{n}{t} - 1)tk, \ldots, r_t, r_t + tk, \ldots, r_t + (\tfrac{n}{t} - 1)tk\}.$$

This design is called a multi-start systematic sampling (MSSS). An unbiased estimator for the sampling variance can be obtained as;

$$\hat{Var}(\bar{y}) = \tfrac{1-f}{t(t-1)} \textstyle\sum_{i=1}^{t} (\bar{y}_i - \bar{y})^2 = \tfrac{k-1}{t(t-1)k} \textstyle\sum_{i=1}^{t} (\bar{y}_i - \bar{y})^2,$$

where, $f = n/N$, $\bar{y} = \frac{1}{t} \sum_{i=1}^{t} \bar{y}_i$ and $\bar{y}_i$ is the subsample mean.

The approach of multiple random starts (MRS) has been incorporated into different systematic sampling methods in order to derive unbiased estimators for the variance of the proposed new estimators of the population mean (see, sec. 2.4.2).

2.3.3 Markov Systematic Sampling

Sampath and Uthayakumaran (1998) introduced a new sampling scheme with Markovian behavior which yields positive inclusion probabilities for all pairs of units. This sampling scheme overcomes the difficulties in Markov sampling proposed by Chandra *et al.* (1991). Markov systematic sampling procedure assumes that the sample size is even and the population size is a multiple of the sample size (i.e. $k = N/n$ is an integer). The population is divided into $n/2$ groups in a systematic manner, say $S_1, \ldots, S_m$ where $m = n/2$ and the i^{th} group (S_i) includes those units with indices $\{2(i-1)k + j; j = 1, 2, \ldots, 2k\}$. For each group S_i, define a transition probability matrix (TPM) A_i of a Markov chain with state space $\{2(i-1)k + j; j = 1, 2, \ldots, 2k\}$. To guarantee that all the sampled units

[2] The general form for the Horvitz-Thompson estimator for the population mean is given by $\bar{y} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$, where π_i is the inclusion probability of the i^{th} unit in the sample (Horvitz and Thompson 1952).

will be distinct, the diagonal elements in each A_i should be zero, $i = 1, 2, \ldots, m$. Selecting a sample of size n using this procedure involves two steps. First, a random number r is drawn from 1 to $2k$. A systematic sample of size $m = n/2$ is then selected using $2k$ as the sampling interval. The units with indices $\{r, r+2k, \ldots, r+2(m-1)k\}$ will be the selected units. A single unit is then drawn from each group independently using the elements of A_i as conditional probabilities. They assumed that the non-diagonal elements of the TPM A_i (a_{rs}) are selected so that for each r;

$$a_{rs} \propto \tau^{|r-s|}, s = 1, 2, \ldots, 2k,$$

where τ is a predetermined positive number which can be chosen either to be the same for all TPMs or to be different τ for the different TPMs. If the same τ is used, we will have a common TPM, say A. The sampling variance can be estimated with the help of the Horvitz-Thompson estimator since all the pairs of units have non-zero inclusion probability.

Kao *et al.* (2011) proposed the remainder Markov systematic sampling design that extends the RLSS and Markov systematic sampling in an attempt to solve the two main statistical problems of the LSS simultaneously. According to their design, selecting a sample of size n involves the following two steps.

1. Divide the population into two strata; the first stratum contains the front $(n-r)k$ units and the second stratum contains the remaining $r(k+1)$ units.
2. Apply the Markov systematic sampling method to each stratum.

2.4 Methods Appropriate for Populations that Exhibit Certain Trend

In literature there are several ways to improve the performance of systematic sampling in the presence of a linear or parabolic trend in the population. One way is to use a weighted mean instead of the unweighted one (usual sample mean $\bar{y}$). On the other hand, one may change the method of sample selection so that the sample mean is not affected by the presence of certain trend. Such two approaches are discussed below in details.

2.4.1 Systematic Sampling with End Corrections

Yates (1948) suggested using a weighted mean in which all internal units of the sample have weight unity (before multiplying by n^{-1}) but the first and last units have different weights. If the random number selected between 1 and k is r, the weights of the first and last unit are

$$1 \pm \frac{n(2r-k-1)}{2(n-1)k}.$$

Using the $+$ sign for the first unit, the $-$ sign for the last, the weighted mean is

$$\bar{y}_w = \bar{y} + \frac{2r-k-1}{2(n-1)k}(y_r - y_{r+(n-1)k})$$

It can be easily verified that if the population consists solely of a linear trend and $N = nk$, the suggested weighted mean coincides with the true population mean.

Bellhouse and Rao (1975) applied the Yates end corrections to the CSS of Lahiri (1951) to improve its performance in the presence of linear trend. Two cases arise while estimating the population mean as follows:

Case 1. The random start taken between 1 and N is small enough, $r + (n-1)k \leq N$. The weights for the first and last unit selected in the sample are

$$1 \pm \frac{n[2r+(n-1)k-(N+1)]}{2(n-1)k}.$$

As before, the $+$ sign is used for the first unit, the $-$ sign for the last.

Case 2. The random start taken between 1 and N is large enough, $r + (n-1)k > N$. Let n_2 be the number of sampled units obtained after passing the N^{th} unit in the population. Then the weights are defined as

$$1 \pm \frac{[2nr+n(n-1)k-n(N+1)-2n_2N]}{2(N-k)}.$$

The $+$ sign is still used for the first unit, the $-$ sign for the last.

2.4.2 Modifying the Method of Sample Selection

Sethi (1965) proposed the balanced systematic sampling (BSS) scheme. With $N = nk$ and n is even, the population is first divided into $n/2$ strata each of size $2k$. Two units equidistant from the end of each stratum are selected in the sample. If the selected random start is $1 \leq r \leq 2k$, the sampled units will have the following indices:

$$[r+2jk,\ \ 2(j+1)k-r+1] \quad ; \qquad j = 0, 1, \ldots, \tfrac{n}{2} - 1$$

With n odd, the sample will be

$$[r+2jk,\ \ 2(j+1)k-r+1,\quad r+(n-1)k] \quad ; \qquad j = 0, 1, \ldots, \tfrac{(n-1)}{2} - 1$$

Singh *et al.* (1968) proposed the modified systematic sampling design (MSS) that can be used for populations exhibiting linear trend. To choose a sample of even size n using this design, we first select a random number r from 1 to k. Then, each pair of units equidistant from the ends of the population is drawn systematically. The sample corresponding to the random start $1 \leq r \leq k$, will contain the units with indices;

$$[r+jk,\quad N-r-jk+1] \quad ; \qquad j = 0, 1, \ldots, \tfrac{n}{2} - 1.$$

With n odd, the sample will be

$$[r+jk, \quad N-r-jk+1, \quad r+\tfrac{1}{2}(n-1)k] \quad ; \qquad j=0,1,\ldots,\tfrac{(n-1)}{2}-1.$$

Madow (1953) introduced a centered systematic sampling scheme in which the random start r is taken as $\frac{k}{2}$ or $\frac{k+2}{2}$ for even $k's$ and $\frac{k+1}{2}$ for odd $k's$. This means that we select the cluster in the center of the possible k clusters.

Sampath and Ammani (2010) applied the MRS approach to the BSS due to Sethi (1965) and the MSS of Singh *et al.* (1968) to provide an unbiased estimator for the sampling variance under these designs. The new designs are described in the following:

I. Balanced Systematic Sampling with Multiple Random Starts (BSSM)

To select a sample of size n using BSSM design with t random starts, one can proceed as follows. First, the population is divided into $n/2t$ groups each of $2tk$ units. Then t random numbers are selected from 1 to tk. Corresponding to every random number chosen, pairs of units equidistant from the group ends are selected in the sample. Clearly, each random number contributes n/t units to the sample. Therefore, the sample will contain n units. Under this design, the sample mean was proved to be an unbiased estimator for the population mean. It was also proved to coincide with the population mean in the presence of linear trend.

II. Modified Systematic Sampling with Multiple Random Starts (MSSM)

When the sample is desired to be selected using MSSM design, then instead of choosing one random start, t random starts are chosen between 1 and tk. Corresponding to every random start selected, pairs of units equidistant from the population ends are selected in the sample in a systematic manner. The sample corresponding to the random start r_i ; $i=1,2,\ldots,t$, will consist of the n/t units with indices;

$$[r_i+jtk, \quad N-r_i-jtk+1] \quad ; \quad j=0,1,\ldots,\tfrac{n}{2t}-1.$$

The population mean is estimated unbiasedly under this design using the sample mean.

2.5 Comparing Different Varieties of Systematic Sampling

In this section comparisons of different versions of systematic sampling are introduced in the sake of highlighting their relative performance.

2.5.1 Comparing Procedures Handling Variable Sample Size Problem

Chang and Huang (2000) assessed the performance of the RLSS relative to SRS, CSS and NPSS under various types of populations and they mentioned the following conclusions:

a. For populations in random order, the RLSS is more efficient than SRS if and only if the sample proportion of the second stratum is larger than the second stratum variance proportion.

b. For populations exhibiting linear trend, $Y_i = i; i = 1, 2, \ldots, N$, the RLSS outperforms SRS if N is not a multiple of the sample size.

c. The RLSS is more efficient than SRS and NPSS for the three types of autocorrelated populations, namely, populations with linear, exponential and hyperbolic correlogram. Also, the RLSS outperforms the CSS for populations with linear correlogram and they are equally efficient for the other types.

2.5.2 Comparing Procedures Providing Unbiased Estimators of the Sampling Variance

Zinger (1980) compared the PSS design with the multi-start systematic sampling. The PSS design is proved to outperform the MSSS if $\rho_w > 0$ and $t = 2$ or 3, or if $\rho_w < 0$ and $t \geq 4$, where t is the number of random starts and ρ_w is the correlation coefficient between pairs of units that are in the same systematic sample.

Leu and Tsui (1996) compared the NPSS with some other sampling procedures under different types of superpopulation. They concluded that:

a. For populations in random order, the NPSS, LSS and SRS are equally efficient.

b. For the auto-correlated populations - linear, exponential and hyperbolic correlogram- NPSS is more efficient than SRS. But NPSS is less efficient than CSS and LSS procedures. In most of the cases, the NPSS outperforms the systematic sampling with multiple random starts.

c. For populations with periodic variation, NPSS is more efficient than LSS. In these populations, the efficiency of the LSS depends on the sampling interval k as mentioned above.

Huang (2004) investigated the performance of the MRSS relative to SRS, CSS and NPSS under different types of population. Huang concluded that:

a. The four designs are equally efficient for populations in random order.

b. For populations with linear trend, $Y_i = \alpha + \beta(i)$, the MRSS is consistently more efficient than SRS, and more efficient than CSS and NPSS in some cases.

c. For the auto-correlated populations, the results are similar to those of the populations with linear trend.

Gautschi (1957) compared the MSSS with the LSS under different types of populations and concluded that the MSSS is more efficient in most cases. Hence, the researcher is

better off choosing MSSS as it provides an unbiased estimator for the sampling variance.

Sampath and Ammani (2010) assessed the performance of both BSSM and MSSM relative to each other and relative to the MSSS under the model that is suitable for populations with linear trend. The model is as follows;

$$Y_i = \alpha + \beta(i) + e_i$$

where, $E(e_i) = 0, V(e_i) = \sigma^2(i)^g$ and $Cov(e_i, e_j) = 0 \quad \forall i \neq j$.

They concluded that:

a. The proposed designs were proved to provide estimators for the population mean that coincide with this mean in the presence of linear trend i.e. they estimate the population mean without any error.

b. For all choices of g and n, BSSM and MSSM are equally efficient as long as we use two random starts.

c. For all choices of g and n, BSSM and MSSM are more efficient than MSSS.

Sampath and Uthayakumaran (1998) compared Markov systematic sampling with SRS, LSS, stratified random sampling and systematic sampling with two random starts for the populations exhibiting exponential trend. For the deterministic exponential model ($Y_i = \beta^i; i = 1, 2, \ldots, N$), Markov systematic sampling has been found to outperform the other procedures. Also, the estimator under Markov systematic sampling is more efficient than the estimators under LSS and systematic sampling with two random starts for populations exhibiting approximate exponential trend ($Y_i = \beta^i + e_i; i = 1, 2, \ldots, N$ and $E(e_i) = 0, E({e_i}^2) = \sigma^2, E(e_i e_j) = 0 \; \forall i \neq j$) when $\tau > 1.2, \beta > 1$ and $\sigma^2 < 75$.

2.5.3 Comparing Systematic Sampling Schemes for Eliminating Trend Effect

Bellhouse and Rao (1975) assessed the performance of both centered systematic sampling, BSS and MSS relative to LSS with Yates corrections and LSS under superpopulation models representing linear and parabolic trends and periodic and autocorrelated variations. In the presence of a linear or parabolic trend, all the three methods, Yates corrections, BSS and MSS, outperform the ordinary LSS.

After reviewing the systematic sampling approach and its derivative designs proposed to handle one or more of the mentioned statistical issues involved in this design, we shall introduce our proposed design in the following chapter. In this chapter, Chapter 3, the proposed design and estimators for the population mean and the sampling variance under this design will be introduced.

3

The Proposed Design and Estimation

The proposed design is a systematic sampling scheme which is a multiple random starts analogue of RLSS. Unbiased estimators for both the population mean and the sampling variance are derived under the proposed design. The following presents this design and the derivations of the estimators in details.

3.1 Remainder Linear Systematic Sampling with Multiple Random Starts (RLSSM)

According to the RLSS design, the population can be divided into two strata; the front $(n-r)k$ units as a stratum and the remaining $r(k+1)$ units as another stratum, as mentioned in Section 2.2.2. Based on this idea, our proposed RLSSM design proceeds as follows:

a. For the first stratum:
Select $1 < t_1 < (n-r)$ different random numbers from the front $t_1 k$ units such that $\frac{(n-r)}{t_1}$ is integer;

$$1 \leq C_1, C_2, \ldots, C_{t_1} \leq t_1 k.$$

Then for each random number chosen select a systematic sample by adding $t_1 k$ as the sampling interval. The sampled units will be;

$$S' = \{y_{C_i+(l'-1)t_1 k} \;\; ; \quad i = 1,2,\ldots,t_1 \;, \quad l' = 1,2,\ldots,\tfrac{(n-r)}{t_1}\}$$

b. For the second stratum:
Select $1 < t_2 < r$ different random numbers from $[(n-r)k+1]$ to $[(n-r)k+t_2(k+1)]$ such that $\frac{r}{t_2}$ is integer;

$$1 \leq C_1, C_2, \ldots, C_{t_2} \leq t_2(k+1)$$

Then for each random number chosen select a systematic sample by adding $t_2(k+1)$ as the sampling interval. The sample will be;

$$S'' = \{y_{(n-r)k+C_i+(l''-1)t_2(k+1)} \;;\; i = 1,2,\ldots,t_2 \;,\; l'' = 1,2,\ldots,\tfrac{r}{t_2}\}$$

The desired sample will be the union of $S'$and $S''$and of size $n = (n-r)+r$.

It is worth noting that the proposed design can be considered as a generalized systematic sampling design. This note is due to the fact that four other designs can be obtained as special cases of the proposed one. Moreover, two special cases of other designs can be obtained as special cases of RLSSM design. Table (1) shows the relations between RLSSM design and the other designs.

Table 1. RLSSM design and its special cases.

N	r	t_1*	t_2	Sampling Design
$nk+r$	$0<r<n$	$1<t_1<(n-r)$	$1<t_2<r$	RLSSM
$nk+r$	$0<r<n$	$t_1=1$	$t_2=1$	RLSS
$nk+r$	$0<r<n$	$t_1=(n-r)$	$t_2=r$	STRS (2 strata)
$nk+r$	$0<r<n$	$t_1=(n-r)$	$t_2=1$	MRSS ($t=1$)
nk	$r=0$	$1<t_1<n$	—	MSSS
nk	$r=0$	$t_1=1$	—	LSS
nk	$r=0$	$t_1=n$	—	SRS

* t_1 and t_2 are selected such that $\frac{(n-r)}{t_1}$ and $\frac{r}{t_2}$ are integers, respectively.

From Table (1), RLSS design is just a RLSSM when $t_1 = t_2 = 1$ i.e. only one random start is taken from each subpopulation. Also, for $r = 0$ our proposed design can produce one of three different designs depending on t_1, the number of random starts. If $1 < t_1 < n$, our design will be reduced to MSSS. On the other hand, if t_1 is one of its two extremes, $t_1 = 1$ or $t_1 = n$, the produced design will be either LSS or SRS, respectively.

Additionally, a special case of MRSS design of Huang (2004) where the random start $t = 1$ can be obtained from RLSSM when $t_1 = (n-r)$ and $t_2 = 1$. In the former case, if $t_2 = r$ instead of 1, the proposed design wil be equivalent to a stratified random sample of size n with only two strata.

3.2 Estimators and their Unbiasedness

Based on the two samples, one can estimate the population mean $\bar{Y}$ as follows:

Step 1. Estimate the first subpopulation mean $\bar{Y}_1$ based on the sample S' using the sample mean in the form,

$$\bar{y}_1 = \tfrac{1}{(n-r)} \textstyle\sum_{i=1}^{t_1} \sum_{l'=1}^{(n-r)/t_1} y_{c_i+(l'-1)t_1 k}$$

$$\bar{y}_1 = \tfrac{1}{(n-r)} \textstyle\sum_{i=1}^{t_1} \tfrac{(n-r)}{t_1} \bar{y}_i$$

where

$$\bar{y}_i = \tfrac{t_1}{(n-r)} \textstyle\sum_{l'=1}^{(n-r)/t_1} y_{c_i+(l'-1)t_1 k}$$

Thus,

$$\bar{y}_1 = \frac{1}{t_1}\sum_{i=1}^{t_1}\bar{y}_i \tag{3.1}$$

Step 2. Estimate the second subpopulation mean $\bar{Y}_2$ based on the sample S'' using the sample mean in the form,

$$\bar{y}_2 = \tfrac{1}{r}\textstyle\sum_{i=1}^{t_2}\sum_{l''=1}^{r/t_2} y_{(n-r)k+c_i+(l''-1)t_2(k+1)}$$

$$\bar{y}_2 = \tfrac{1}{r}\textstyle\sum_{i=1}^{t_2}\frac{r}{t_2}\bar{y}_i$$

where

$$\bar{y}_i = \tfrac{t_2}{r}\textstyle\sum_{l''=1}^{r/t_2} y_{(n-r)k+c_i+(l''-1)t_2(k+1)}$$

Thus,

$$\bar{y}_2 = \frac{1}{t_2}\sum_{i=1}^{t_2}\bar{y}_i \tag{3.2}$$

Step 3. The estimator of the population mean can be taken as a weighted mean of the two means stated above in (3.1) and (3.2):

$$\bar{y}_{RLSSM} = \frac{(n-r)k\bar{y}_1 + r(k+1)\bar{y}_2}{N} \tag{3.3}$$

Lemma 1: Under this design (RLSSM) the first and second order inclusion probabilities for units in the first subpopulation [1 to $(n-r)k$] are defined as follows:

$$\pi_i = \tfrac{1}{k} \quad \forall\, i = 1,2,\ldots,(n-r)k$$

$$\pi_{ij} = \tfrac{1}{k} \quad \forall \quad i,j \in S_h, \quad i \neq j, \quad h = 1,2,\ldots,t_1$$

$$\pi_{ij} = \tfrac{t_1-1}{k(t_1k-1)} \quad \forall \quad i \in S_h, \quad j \in S_l, \quad h \neq l, \quad h,l = 1,2,\ldots,t_1$$

Proof: Let us look at the t_1-start systematic sample chosen from the first subpopulation from another side following Sampath (2012). To choose a sample of size $(n-r)$ from that population, it is first divided into t_1k groups of $(n-r)/t_1$ units each, as in Table (2), and t_1 of these groups will be randomly selected to get the desired sample.

Table 2. Partitioning the first subpopulation under MSSS.

	Units					
	S_1	1	t_1k+1	$2t_1k+1$	...	$[(n-r)/t_1-1]t_1k+1$
	S_2	2	t_1k+2	$2t_1k+2$	...	$[(n-r)/t_1-1]t_1k+2$
	...	...	...	...	...	...
	...	...	...	...	...	...
Groups	S_i	i	t_1k+i	$2t_1k+i$	...	$[(n-r)/t_1-1]t_1k+i$
	...	...	...	...	...	...
	...	...	...	...	...	...
	S_{t_1k}	t_1k	$2t_1k$	$3t_1k$	...	$(n-r)k$

The probability of including the unit with label i in the sample (π_i) is exactly the probability of including the group containing this unit in the sample. Therefore, the first order inclusion probabilities can be defined as follows:

$$\pi_i = \frac{t_1}{t_1 k} = \frac{1}{k} \quad \forall\, i = 1,2,\ldots,(n-r)k$$

Following the same view, it should be noticed that the second order inclusion probabilities for pairs of units which belong to the same group are equal to the probability of selecting this group in the sample which is equal to the first order inclusion probabilities.

On the other hand, if the pair of units belongs to two different groups, then including this pair in the sample can be only realized by selecting the two groups in the sample. Hence, the second order inclusion probability in this case will be:

$$\pi_{ij} = \frac{t_1}{t_1 k} * \frac{t_1-1}{t_1 k-1} = \frac{t_1-1}{k(t_1 k-1)} \quad \forall \quad i \in S_h, \quad j \in S_l, \quad h \neq l, \quad h,l = 1,2,\ldots,t_1$$

The proof is complete.

Lemma 2: Under this design (RLSSM) the first and second order inclusion probabilities for units in the second subpopulation [$(n-r)k+1$ to N] are defined as follows:

$$\pi_i = \frac{1}{k+1} \quad \forall\, i = (n-r)k+1,\ldots,N$$

$$\pi_{ij} = \frac{1}{k+1} \quad \forall \quad i,j \in S_g, \quad i \neq j, \quad g = 1,2,\ldots,t_2$$

$$\pi_{ij} = \frac{t_2-1}{(k+1)[t_2(k+1)-1]} \quad \forall \quad i \in S_g, \quad j \in S_p, \quad g \neq p, \quad g,p = 1,2,\ldots,t_2$$

Proof: To choose a sample of size r from the second subpopulation, this population is first divided into $t_2(k+1)$ groups of (r/t_2) units each, as in Table (3), and t_2 of these groups will be randomly selected to get the desired sample.

Table 3. Partitioning the second subpopulation under MSSS.

	Units					
Groups	S_1	1	$t_2(k+1)+1$	$2t_2(k+1)+1$	...	$[r/t_2-1]t_2(k+1)+1$
	S_2	2	$t_2(k+1)+2$	$2t_2(k+1)+2$	...	$[r/t_2-1]t_2(k+1)+2$
	...	...	...	...	...	...
	...	...	...	...	...	...
	S_i	i	$t_2(k+1)+i$	$2t_2(k+1)+i$	...	$[r/t_2-1]t_2(k+1)+i$
	...	...	...	...	...	...
	...	...	...	...	...	...
	$S_{t_2(k+1)}$	$t_2(k+1)$	$2t_2(k+1)$	$3t_2(k+1)$	...	$r(k+1)$

The rest of the proof follows the same procedure used in the proof of Lemma 1.

Theorem 1: The sample mean of the RLSSM design, $\bar{y}_{RLSSM}$, is an unbiased estimator for the population mean $\bar{Y}$.

Proof: Since, $\bar{y}_1$ and $\bar{y}_2$ are unbiased estimators for the subpopulation means, $\bar{Y}_1$ and $\bar{Y}_2$, respectively, (Gautschi (1957)), then

$$E(\bar{y}_{RLSSM}) = \frac{(n-r)k\bar{Y}_1 + r(k+1)\bar{Y}_2}{N}$$

$$= \frac{\sum_{j=1}^{(n-r)k} Y_j + \sum_{j=(n-r)k+1}^{N} Y_j}{N}$$

$$= \frac{1}{N}\sum_{j=1}^{N} Y_j = \bar{Y}.$$

This proves the theorem.

Remark 1: It can be easily shown that the proposed estimator $(\bar{y}_{RLSSM})$ is in fact a Horvitz-Thomson (1952) estimator in the form $\hat{\bar{Y}}_{HT} = \frac{1}{N}\sum_{u_i \in S} \frac{y_i}{\pi_i}$ as follows:

$$\bar{y}_{RLSSM} = \frac{1}{N}[(n-r)k.\frac{1}{n-r}\sum_{i=1}^{t_1}\sum_{l'=1}^{\frac{n-r}{t_1}} y_{c_i+(l'-1)t_1 k}$$

$$+ r(k+1).\frac{1}{r}\sum_{i=1}^{t_2}\sum_{l''=1}^{\frac{r}{t_2}} y_{(n-r)k+c_i+(l''-1)t_2(k+1)}]$$

$$= \frac{1}{N}[\sum_{i=1}^{t_1}\sum_{l'=1}^{\frac{n-r}{t_1}} \frac{y_{c_i+(l'-1)t_1 k}}{1/k} + \sum_{i=1}^{t_2}\sum_{l''=1}^{\frac{r}{t_2}} \frac{y_{(n-r)k+c_i+(l''-1)t_2(k+1)}}{1/(k+1)}]$$

$$= \frac{1}{N}[\sum_{U_i \in S'} \frac{y_i}{\pi_i} + \sum_{U_j \in S''} \frac{y_j}{\pi_j}].$$

Therefore, an unbiased estimator for the sampling variance can be derived based on Yates - Grundy (1953) estimator which has the following form;

$$\hat{Var}(\bar{y}_n) = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$$

However, the terms $(\pi_i\pi_j - \pi_{ij})$ are sometimes negative under the proposed design and so $\hat{Var}(\bar{y}_n)$ may be sometimes negative. Thus the sampling variance of this design and another unbiased estimator that is always positive are given by the following two theorems.

Theorem 2: Under the RLSSM design, the variance of the sample mean has the form:

$$Var(\bar{y}_{RLSSM}) = \frac{1}{N^2}\{\frac{(n-r)^2 k(k-1)}{t_1(t_1 k-1)}\sum_{i=1}^{t_1 k}(\bar{Y}_{1i} - \bar{Y}_1)^2 + \frac{r^2 k(k+1)}{t_2[t_2(k+1)-1]}\sum_{i=1}^{t_2(k+1)}(\bar{Y}_{2i} - \bar{Y}_2)^2\} \tag{3.4}$$

Proof: Since, the random starts are chosen independently, $\bar{y}_1$ and $\bar{y}_2$ are independent. Thus $Var(\bar{y}_{RLSSM})$ can be written as follows;

$$Var(\bar{y}_{RLSSM}) = \frac{1}{N^2}[(n-r)^2k^2.Var(\bar{y}_1) + r^2(k+1)^2.Var(\bar{y}_2)]$$

$$= \frac{1}{N^2}\{(n-r)^2k^2(1-\frac{1}{k})\frac{{s_1}^2}{t_1} + r^2(k+1)^2(1-\frac{1}{k+1})\frac{{s_2}^2}{t_2}\} \qquad (3.5)$$

where, ${S_1}^2 = \frac{1}{t_1k-1}\sum_{i=1}^{t_1k}(\bar{Y}_{1i}-\bar{Y}_1)^2$ is the variance between clusters' means in the first stratum and ${S_2}^2 = \frac{1}{t_2(k+1)-1}\sum_{i=1}^{t_2(k+1)}(\bar{Y}_{2i}-\bar{Y}_2)^2$ is the variance between clusters' means in the second stratum.

Substituting these quantities in (3.5), the theorem follows.

Theorem 3: The sampling variance of the RLSSM design can be estimated unbiasedly as follows:

$$\hat{Var}(\bar{y}_{RLSSM}) = \frac{1}{N^2}\{\frac{(n-r)^2k(k-1)}{t_1(t_1-1)}\sum_{j=1}^{t_1}(\bar{y}_{1j}-\bar{y}_1)^2 + \frac{r^2k(k+1)}{t_2(t_2-1)}\sum_{j=1}^{t_2}(\bar{y}_{2j}-\bar{y}_2)^2\} \quad (3.6)$$

where, $\bar{y}_{1j}\ ; j = 1,\ldots,t_1$ is the mean of the j^{th} sample from the first subpopulation and $\bar{y}_{2j}\ ; j = 1,\ldots,t_2$ is the mean of the j^{th} sample from the second subpopulation.

Proof: Since it can be proved that $\hat{{S_1}^2} = \frac{1}{t_1-1}\sum_{j=1}^{t_1}(\bar{y}_{1j}-\bar{y}_1)^2$ and $\hat{{S_2}^2} = \frac{1}{t_2-1}\sum_{j=1}^{t_2}(\bar{y}_{2j}-\bar{y}_2)^2$ are unbiased estimators for ${S_1}^2$ and ${S_2}^2$ respectively, as follows;

$$E(\hat{{S_1}^2}) = \frac{1}{t_1-1}E[\sum_{j=1}^{t_1}(\bar{y}_{1j}-\bar{y}_1)^2]$$

$$= \frac{1}{t_1-1}E\sum_{j=1}^{t_1}[(\bar{y}_{1j}-\bar{Y}_1)^2 + (\bar{y}_1-\bar{Y}_1)^2 - 2(\bar{y}_{1j}-\bar{Y}_1)(\bar{y}_1-\bar{Y}_1)]$$

$$= \frac{1}{t_1-1}E[\sum_{j=1}^{t_1}(\bar{y}_{1j}-\bar{Y}_1)^2 - t_1(\bar{y}_1-\bar{Y}_1)^2]$$

$$= \frac{1}{t_1-1}[t_1(\frac{1}{t_1k})\sum_{i=1}^{t_1k}(\bar{Y}_{1i}-\bar{Y}_1)^2 - t_1(\frac{k-1}{t_1k})(\frac{1}{t_1k-1})\sum_{i=1}^{t_1k}(\bar{Y}_{1i}-\bar{Y}_1)^2]$$

$$= \frac{1}{t_1-1}[\frac{1}{k}\sum_{i=1}^{t_1k}(\bar{Y}_{1i}-\bar{Y}_1)^2(1-\frac{k-1}{t_1k-1})] = \frac{1}{t_1k-1}\sum_{i=1}^{t_1k}(\bar{Y}_{1i}-\bar{Y}_1)^2 = {S_1}^2$$

Using similar procedure, $\hat{{S_2}^2}$ is an unbiased estimator for ${S_2}^2$. This completes the theorem.

3.3 A Numerical Illustration for RLSSM Procedure

Consider the population of size 40 units given in Cochran (1977, p.211, Table 8.3) which is an artificial population that exhibits a fairly steady rising trend. If the population mean is needed to be estimated on the basis of a sample of size 12 units, using RLSSM one can proceed as follows:

Table 4. Units of an artificial fairly steady rising population.

0	1	1	2	5	4	7	7	8	6
6	8	9	10	13	12	15	16	16	17
18	19	20	20	24	23	25	29	29	27
26	30	31	31	33	32	35	37	38	38

$$N = 40, \quad \bar{Y} = 18.175, \quad n = 12, \quad \frac{N}{n} = 3.33, \quad k = 3$$
$$N = (12*3)+4 = (n*k)+r \quad ; r = 4$$
$$N = (n-r)k + r(k+1) = (8*3)+(4*4) = 40$$

From the first subpopulation, $1,2,\ldots,(n-r)k = 24$, select a multi-start systematic sample of size 8 units using $t_1 = 2$ random starts. Let the first random start chosen from $1,2,\ldots,t_1 k = 6$, be $C_1 = 4$ then,

$$S_1{}' = \{y_4, y_{10}, y_{16}, y_{22}\}.$$

If $C_2 = 2$,

$$S_2{}' = \{y_2, y_8, y_{14}, y_{20}\}$$

and the sample from the first subpopulation will be:

$$S' = \{y_2, y_4, y_8, y_{10}, y_{14}, y_{16}, y_{20}, y_{22}\} = \{1,2,7,6,10,12,17,19\}.$$

Thus,

$$\bar{y}_1 = \frac{9.75+8.75}{2} = 9.25.$$

From the second subpopulation, $(n-r)k+1 = 25,\ldots,N = 40$, select a multi-start systematic sample of size 4 units using $t_2 = 2$ random starts. Let the first random start chosen from $25,26,\ldots,(n-r)k+t_2(k+1) = 32$, be $C_1 = 29$ then,

$$S_1{}'' = \{y_{29}, y_{37}\}.$$

If $C_2 = 26$,

$$S_2{}'' = \{y_{26}, y_{34}\}$$

and the sample from the second subpopulation will be:

$$S'' = \{y_{26}, y_{29}, y_{34}, y_{37}\} = \{23,29,31,35\}.$$

$$\bar{y}_2 = \frac{32+27}{2} = 29.5$$

$$\bar{y}_{RLSSM} = 17.35$$

$$\hat{Var}(\bar{y}_{RLSSM}) = 0.81$$

If these 12 units are selected randomly, SRS;

$$\bar{y}_{SRS} = \frac{192}{12} = 16$$

$$\hat{Var}(\bar{y}_{SRS}) = 7.254$$

So, for this type of populations, RLSSM is highly more efficient than SRS.

4

Performance Comparisons

Efficiencies of systematic sampling designs depend on the characters of the sampled populations. Thus, efficiencies of these sampling procedures are compared for various types of populations. Instead of considering a single finite population, $y_1, \ldots, y_N$, it will be assumed, following Cochran (1946), that the $y_i's$ are drawn from an infinite superpopulation having some specified properties. Hence, the performance comparisons will be carried out on the basis of comparing the expected variances, where the expectation is taken over the assumed superpopulation model, rather than the variances directly.

More specifically, the comparisons will be done under three types of superpopulations, namely, populations in random order, populations with linear trend and autocorrelated populations. Under each of these populations, the performance of the proposed design, RLSSM, will be assessed relative to five other sampling designs, namely, SRS, CSS, RLSS, NPSS and MRSS. Both of CSS and RLSS can handle the problem of non-integer sampling intervals (k) but they do not provide an unbiased estimator for the sampling variance. On the other hand, each of NPSS and MRSS can tackle the two main problems of the linear systematic sampling (LSS) simultaneously.

4.1 Populations in Random Order

Under this model, according to Cochran (1977), the variates $y_i(1,2,\ldots,N)$ are assumed to be uncorrelated having the same expectations while the variances may change with i,

$$E(y_i) = \mu, \quad E(y_i - \mu)^2 = {\sigma_i}^2, \quad i = 1,2,\ldots,N, \quad E(y_i - \mu)(y_j - \mu) = 0 \quad \forall i \neq j. \tag{4.1}$$

Under model (4.1), the expected variance of SRS is given in Cochran (1977) by,

$${\sigma^2}_{SRS} = (\frac{1}{n} - \frac{1}{N})\sigma^2 \qquad ; \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} {\sigma_i}^2 \tag{4.2}$$

It is worth noting that formula (4.2) holds for any sampling design with fixed sample size n and identical first order inclusion probability for all units. Thus the expected variances of CSS and NPSS and MRSS are proved by Leu and Tsui (1996) and Huang (2004), respectively, to be equal to (4.2).

Theorem 4: Under the model for randomly ordered populations, given by (4.1), the expected variance of RLSSM is

$$\sigma^2{}_{RLSSM} = \frac{1}{N^2}\{(k-1)\sum_{i=1}^{(n-r)k}\sigma_i{}^2 + k\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2\} \tag{4.3}$$

Proof: Based on the fact in (4.4), given by Gautschi (1957), which relates the sampling variance of MSSS with that of LSS, the variance of our proposed RLSSM design can be derived as follows.

$$Var(\bar{y}_{MSSS}) = (\frac{k-1}{tk-1})\frac{1}{tk}\sum_{i=1}^{tk}(\bar{y}_i - \bar{Y})^2 = (\frac{k-1}{tk-1})Var^{(n/t)}(\bar{y}_{LSS}) \tag{4.4}$$

where, $Var^{(n/t)}(\bar{y}_{LSS})$ is the sampling variance of a LSS of size (n/t).

Taking the expectation of (3.4) with respect to model (4.1) gives

$$\sigma^2{}_{RLSSM} = \tfrac{1}{N^2}\{\tfrac{(n-r)^2k^2(k-1)}{(t_1k-1)}[\tfrac{(t_1k-1)}{t_1k}\textstyle\sum_{i=1}^{(n-r)k}\tfrac{t_1\sigma_i{}^2}{(n-r)^2k}] + \tfrac{r^2k(k+1)^2}{t_2(k+1)-1}[\tfrac{t_2(k+1)-1}{t_2(k+1)}\textstyle\sum_{i=(n-r)k+1}^{N}\tfrac{t_2\sigma_i{}^2}{r^2(k+1)}]\}$$

$$= \tfrac{1}{N^2}\{(k-1)\textstyle\sum_{i=1}^{(n-r)k}\sigma_i{}^2 + k\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2\}$$

Note that in this case $\sigma^2{}_{RLSSM} = \sigma^2{}_{RLSS}$ (see, Chang and Huang (2000)). Hence, our proposed design has the same efficiency as the RLSS design under populations in random order.

Putting $r = 0$ in (4.4), $\sigma^2{}_{RLSSM}$ will be reduced to that of LSS ($\sigma^2{}_{LSS}$), given in Cochran (1977), with sample of size n and $k = N/n$.

$$\sigma^2{}_{LSS} = (\frac{k-1}{k})\frac{\sigma^2}{n} \tag{4.5}$$

Since each of CSS, NPSS and MRSS has the same expected variance as that of SRS given by (4.2), our proposed design will be compared to SRS and the result will be the same for the other three designs.

$$\sigma^2{}_{SRS} - \sigma^2{}_{RLSSM} = (\tfrac{1}{n} - \tfrac{1}{N})\sigma^2 - \tfrac{1}{N^2}\{(k-1)\textstyle\sum_{i=1}^{(n-r)k}\sigma_i{}^2 + k\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2\}$$

$$= \tfrac{N-n}{nN^2}\textstyle\sum_{i=1}^{N}\sigma_i{}^2 - \tfrac{1}{N^2}\{(k-1)\sum_{i=1}^{(n-r)k}\sigma_i{}^2 + k\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2\}$$

$$= \tfrac{(nk+r-n)}{nN^2}\textstyle\sum_{i=1}^{N}\sigma_i{}^2 - \tfrac{n(k-1)}{nN^2}\sum_{i=1}^{(n-r)k}\sigma_i{}^2 - \tfrac{nk}{nN^2}\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2$$

$$= \tfrac{1}{nN^2}\{n(k-1)\textstyle\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2 - nk\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2 + r\sum_{i=1}^{N}\sigma_i{}^2\}$$

$$= \tfrac{1}{nN^2}\{r\textstyle\sum_{i=1}^{N}\sigma_i{}^2 - n\sum_{i=(n-r)k+1}^{N}\sigma_i{}^2\}$$

The RLSSM will be more efficient than SRS if

$$\sigma^2{}_{SRS} - \sigma^2{}_{RLSSM} > 0,$$

or

$$r\sum_{i=1}^{N} \sigma_i{}^2 - n\sum_{i=(n-r)k+1}^{N} \sigma_i{}^2 > 0,$$

or

$$\frac{r}{n} > \frac{\sum_{i=(n-r)k+1}^{N} \sigma_i{}^2}{\sum_{i=1}^{N} \sigma_i{}^2}.$$

This means that the proportion of the sample from the second subpopulation should be greater than the proportion of the second subpopulation variance. So under the given model in (4.1), RLSSM is recommended over each of SRS, CSS, NPSS and MRSS designs if the previous condition holds, cf. Chang and Huang (2000).

4.2 Populations with Linear Trend

If the population consists solely of a linear trend, the variates $y_i's$ are assumed to be equal to the corresponding labels as follows:

$$y_i = i, \quad i = 1,2,\ldots,N \tag{4.6}$$

$$\bar{Y} = \tfrac{N+1}{2} \qquad \text{and} \qquad S^2 = \tfrac{N(N+1)}{12}$$

This type of populations, given by the model in (4.6), is found in Cochran (1977). For these populations, Cochran (1977) showed that

$$Var(\bar{y}_{SRS}) = \frac{(N-n)(N+1)}{12n}. \tag{4.7}$$

Chang and Huang (2000) gave the sampling variance of RLSS in the form:

$$Var(\bar{y}_{RLSS}) = \frac{k}{12N^2}[(n-r)^2k(k^2-1) + r^2(k+1)^2(k+2)] \tag{4.8}$$

Theorem 5, below, generalizes (4.8) into our RLSSM general design.

Theorem 5: Under the model given in (4.6), the sampling Variance of RLSSM is given by:

$$Var(\bar{y}_{RLSSM}) = \frac{k}{12N^2}\{(n-r)^2k(k-1)(t_1k+1) + r^2(k+1)^2[t_2(k+1)+1]\} \tag{4.9}$$

Proof: The sampling variance of LSS is showed by Cochran (1977) to be

$$Var(\bar{y}_{LSS}) = \frac{(k^2-1)}{12} \tag{4.10}$$

Combining (4.4) and (4.10) together we have the following:

$$Var(\bar{y}_{MSSS}) = (\frac{k-1}{tk-1})(\frac{t^2k^2-1}{12}) = \frac{(k-1)(tk+1)}{12}$$

Applying this result for each of the two sub-samples in RLSSM design, the sampling variance of the proposed design can be obtained as follows.

$$\begin{aligned} Var(\bar{y}_{RLSSM}) &= \frac{1}{N^2}\{(n-r)^2k^2(\frac{k-1}{t_1k-1})(\frac{t_1^2k^2-1}{12}) + r^2(k+1)^2(\frac{k}{t_2(k+1)-1})(\frac{t_2^2(k+1)^2-1}{12}\} \\ &= \frac{1}{12N^2}\{(n-r)^2k^2(k-1)(t_1k+1) + r^2(k+1)^2k[t_2(k+1)+1]\} \\ &= \frac{k}{12N^2}\{(n-r)^2k(k-1)(t_1k+1) + r^2(k+1)^2[t_2(k+1)+1]\} \end{aligned}$$

It's worth noting that when $t_1 = t_2 = 1$, equation (4.9) will be reduced to (4.8), the variance of the RLSS. Also, if $r = 0$, (4.9) will be reduced to the variance of the MSSS. In the simplest case, the variance of LSS will be obtained when setting $t_1 = t_2 = 1$ and $r = 0$.

Comparing the sampling variances given by (4.7) and (4.9) indicates the superiority of RLSSM over SRS in terms of efficiency. The equality occurs when $r = 0, t_1 = 1$ and $n = 1$ where the two variances will be reduced to $(k^2-1)/12$.

The sampling variances of CSS, NPSS and MRSS cannot be obtained in a simple form due to the circular nature of these designs. Hence, the comparison with these designs will be carried out numerically.

Table (5) shows the values of sampling variances for SRS, RLSS, CSS, NPSS, MRSS and RLSSM obtained from an emperical study for some chosen values of N and n with the help of the statistical package R.

Table 5. Variances of the sample mean corresponding to six sampling procedures, namely, SRS, RLSS, CSS, NPSS, MRSS and RLSSM.

N	n	u,a	t_1,t_2	σ^2_{SRS}	σ^2_{RLSS}	σ^2_{CSS}	σ^2_{NPSS}	σ^2_{MRSS}	σ^2_{RLSSM}
20	8	12, 4	2, 2	2.6250	0.2800	1.7500	1.8542	1.5625	0.4867
36	8	6, 2	2, 2	10.7917	0.8642	2.9167	3.5104	5.8542	1.5761
36	8	16, 4	2, 2	10.7917	0.8642	2.9167	5.8542	5.8542	1.5761
36	10	4, 2	2, 2	8.0167	0.6296	2.3167	2.3100	3.3567	1.1296
36	10	4, 2	2, 3	8.0167	0.6296	2.3167	2.3100	3.3567	1.5741
40	12	8, 4	2, 2	7.9722	0.4400	2.1667	4.1944	5.6944	0.7800
40	12	8, 4	4, 2	7.9722	0.4400	2.1667	4.1944	5.6944	1.1400
42	12	10, 4	2, 2	8.9583	0.5306	3.9167	3.5972	4.6875	0.9490
42	12	10, 4	3, 3	8.9583	0.5306	3.9167	3.5972	4.6875	1.3673
46	10	6, 2	2, 2	14.1000	1.0019	3.0000	3.5133	5.6800	1.8318
46	10	6, 2	2, 3	14.1000	1.0019	3.0000	3.5133	5.6800	2.5406
48	14	8, 4	2, 2	9.9167	0.4792	3.8810	6.4728	5.9982	0.8542
48	14	8, 4	4, 3	9.9167	0.4792	3.8810	6.4728	5.9982	1.3542
50	14	10, 4	2, 2	10.9286	0.5984	3.4337	3.7092	4.7449	1.0728
50	14	10, 4	2, 4	10.9286	0.5984	3.4337	3.7092	4.7449	1.8920

It is obvious from Table (5) that both RLSS and RLSSM outperform the other designs under this type of populations whatever the number of random starts. The RLSS procedure has higher efficiency than the proposed design but, as noted earlier, does not offer an unbiased estimator for the sampling variance.

4.3 Auto-correlated Populations

According to Cochran (1946), under this kind of populations it is assumed that two elements y_i, y_j are positively correlated with a correlation which depends only on the distance "$d = |j-i|$" and decreases as d increases. The mean and variance of $y_i's$ are supposed to be constant. This type of populations is frequently observed in extensive samplings where the variance within a group of elements increases steadily as the size of the group increases.

$$E(y_i) = \mu, \qquad Var(y_i) = \sigma^2, \quad Cov(y_i, y_j) = \rho_d \sigma^2 \quad \forall i \neq j \tag{4.11}$$

Under this type of populations, the expected variance of SRS was obtained by Cochran (1946) to be as follows.

$$\sigma^2{}_{SRS} = (\frac{1}{n} - \frac{1}{N})\sigma^2[1 - \frac{2}{N(N-1)} \sum_{d=1}^{N-1} (N-d)\rho_d] \tag{4.12}$$

The expected variance of the multi-start systematic sample mean is given by Gautschi (1957) as:

$$\sigma^2{}_{MSSS} = \frac{k-1}{N}\sigma^2[1 - \frac{2}{N(kt-1)} \sum_{d=1}^{N-1} (N-d)\rho_d + \frac{2kt^2}{n(kt-1)} \sum_{d=1}^{\frac{n}{t}-1} (\frac{n}{t} - d)\rho_{ktd}] \tag{4.13}$$

Chang and Huang (2000) gave the expected variance of RLSS in the form:

$$\begin{aligned} \sigma^2{}_{RLSS} = \tfrac{\sigma^2}{N^2}\{ & k(N-n+r) \\ & -2\textstyle\sum_{d=1}^{(n-r)k-1}[(n-r)k-d]\rho_d + 2k^2 \sum_{d=1}^{n-r-1}(n-r-d)\rho_{dk} \\ & -2 \sum_{d=1}^{r(k+1)-1} [r(k+1)-d]\rho_d + 2(k+1)^2 \sum_{d=1}^{r-1}(r-d)\rho_{d(k+1)}\} \end{aligned} \tag{4.14}$$

The expected variances of CSS and MRSS are given in Huang (2004) by:

$$\sigma^2{}_{CSS} = (\frac{1}{n} - \frac{1}{N})\sigma^2 + \frac{2\sigma^2}{Nn^2} \sum_{t=1}^{N} \sum_{i=0}^{n-1} \sum_{j>i}^{n-1} \rho_{|(ik+t)-(jk+t)|} - \frac{2\sigma^2}{N^2} \sum_{d=1}^{N-1} (N-d)\rho_d \tag{4.15}$$

and

$$\begin{aligned} \sigma^2{}_{MRSS} = (\tfrac{1}{n} - \tfrac{1}{N})\sigma^2 + \tfrac{2\sigma^2}{Nn^2} \textstyle\sum_{t=1}^{N} \{ & \tfrac{n-r-1}{k[(n-r)k-1]} \sum_{i=0}^{(n-r)k-1} \sum_{j>i}^{(n-r)k-1} \rho_{|(t+j)-(t+i)|} \\ & + \tfrac{1}{k} \sum_{i=0}^{(n-r)k-1} \sum_{j=1}^{r} \rho_{|(t+i)-[j(k+1)+t+(n-r)k-1]|} \\ & + \textstyle\sum_{i=0}^{r} \sum_{j>i}^{r} \rho_{|[i(k+1)+t+(n-r)k-1]-[j(k+1)+t+(n-r)k-1]|} \} \\ & - \frac{2\sigma^2}{N^2} \sum_{d=1}^{N-1} (N-d)\rho_d \end{aligned} \tag{4.16}$$

The expected variance of the NPSS design is obtained by Leu and Tsui (1996) as follows:

$$\sigma^2{}_{NPSS} = (\tfrac{1}{n} - \tfrac{1}{N})\sigma^2 + \tfrac{2}{Nn^2}\sigma^2\{\tfrac{a(a-1)}{u(u-1)}\textstyle\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{(j>i}^{u-1}\rho_{|(t+j)-(t+i)|}]$$
$$+\tfrac{a}{u}\textstyle\sum_{t=1}^{N}[\sum_{i=0}^{u-1}\sum_{j=1}^{n-a}\rho_{|(t+i)-(jk+t+u-1)|}]$$
$$+\textstyle\sum_{t=1}^{N}[\sum_{i=0}^{n-a}\sum_{j>i}^{n-a}\rho_{|(ik+t+u-1)-(jk+t+u-1)|}]\}$$
$$-\frac{2}{N^2}\sigma^2\sum_{d=1}^{N-1}(N-d)\rho_d \tag{4.17}$$

The expected variance under the RLSSM design is given in the next theorem.

Theorem 6: The expected variance of the mean of the RLSSM design is obtained in the following form:

$$\sigma^2{}_{RLSSM} = \tfrac{\sigma^2}{N^2}\{k(N-n+r) - \tfrac{2(k-1)}{t_1k-1}\textstyle\sum_{d=1}^{(n-r)k-1}[(n-r)k-d]\rho_d + \tfrac{2t_2^2k(k+1)^2}{[t_2(k+1)-1]}$$
$$\times\textstyle\sum_{d=1}^{\frac{r}{t_2}-1}(\tfrac{r}{t_2}-d)\rho_{dt_2(k+1)} + \tfrac{2t_1^2k^2(k-1)}{t_1k-1}\textstyle\sum_{d=1}^{\frac{n-r}{t_1}-1}(\tfrac{n-r}{t_1}-d)\rho_{dt_1k}$$

$$-\frac{2k}{[t_2(k+1)-1]}\sum_{d=1}^{r(k+1)-1}[r(k+1)-d]\rho_d\} \tag{4.18}$$

Proof: Let $E_M[Var(\bar{y}_{RLSSM})] = \xi[Var(\bar{y}_{RLSSM})] = \sigma^2_{RLSSM}$, then we have the following;

$$\sigma^2{}_{RLSSM} = \tfrac{1}{N^2}\{\tfrac{(n-r)^2k^2(k-1)}{t_1k-1}\xi[\tfrac{1}{t_1k}\textstyle\sum_{i=1}^{t_1k}(\bar{y}_{1i}-\bar{Y}_1)^2] + \tfrac{r^2(k+1)^2k}{[t_2(k+1)-1]}$$
$$\times\xi[\tfrac{1}{t_2(k+1)}\textstyle\sum_{i=1}^{t_2(k+1)}(\bar{y}_{2i}-\bar{Y}_2)^2]\}.$$

Let $I = (n-r)^2k^2\frac{k-1}{t_1k-1}\xi[\frac{1}{t_1k}\sum_{i=1}^{t_1k}(\bar{y}_{1i}-\bar{Y}_1)^2]$
and $II = r^2(k+1)^2\frac{k}{[t_2(k+1)-1]}\xi[\frac{1}{t_2(k+1)}\sum_{i=1}^{t_2(k+1)}(\bar{y}_{2i}-\bar{Y}_2)^2]$.

Working on I gives;

$$I = (n-r)^2k^2(\tfrac{k-1}{t_1k-1})\tfrac{t_1\sigma^2}{n-r}(1-\tfrac{1}{t_1k})\{1 - \tfrac{2}{(n-r)k(t_1k-1)}\textstyle\sum_{d=1}^{(n-r)k-1}[(n-r)k-d]\rho_d$$
$$+\tfrac{2t_1^2k}{(n-r)(t_1k-1)}\textstyle\sum_{d=1}^{\frac{(n-r)}{t_1}-1}[\tfrac{n-r}{t_1}-d]\rho_{dt_1k}\}$$
$$= (n-r)k(k-1)\sigma^2\{1 - \tfrac{2}{(n-r)k(t_1k-1)}\textstyle\sum_{d=1}^{(n-r)k-1}[(n-r)k-d]\rho_d + \tfrac{2t_1^2k}{(n-r)(t_1k-1)}$$
$$\times\textstyle\sum_{d=1}^{\frac{n-r}{t_1}-1}[\tfrac{n-r}{t_1}-d]\rho_{dt_1k}\}$$

Working on II in the same fashion one can easily show that;

$$II = r^2(k+1)^2 \frac{k}{[t_2(k+1)-1]} . \frac{t_2\sigma^2}{r}(1-\frac{1}{t_2(k+1)})\{1-\frac{2}{r(k+1)[t_2(k+1)-1]}\sum_{d=1}^{r(k+1)-1}[r(k+1)-d]\rho_d$$

$$+\frac{2t_2^2(k+1)}{r[t_2(k+1)-1]}\sum_{d=1}^{\frac{r}{t_2}-1}[\frac{r}{t_2}-d]\rho_{dt_1(k+1)}\}$$

$$= rk(k+1)\sigma^2\{1-\frac{2}{r(k+1)[(t_2(k+1)-1]}\sum_{d=1}^{r(k+1)-1}[r(k+1)-d]\rho_d + \frac{2t_2^2(k+1)}{r[t_2(k+1)-1]}$$

$$\times\sum_{d=1}^{\frac{r}{t_2}-1}[\frac{r}{t_2}-d]\rho_{dt_1(k+1)}\}$$

Thus, $\sigma^2{}_{RLSSM} = \frac{1}{N^2}[I+II] = (4.18)$. Hence the theorem follows.

It can be easily verified that when $t_1 = t_2 = 1$, expression (4.18) reduces to (4.14). In other words, the expected variance of RLSSM is reduced to that of RLSS when only one random start is selected. Additionaly, if there is no remainder term, $r = 0$, the proposed design is equivalent to the multi-start systematic sampling design and (4.18) reduces to (4.13).

Looking at formulas (4.12) through (4.18) shows the difficulty in obtaining a general result about the relative efficiency of the considered sampling designs. However, performance comparisons can be carried out empirically for three types of correlograms considered by Cochran (1946), which are

i. Linear correlogram: $\rho_d = 1 - d/L \quad ; L \geq N-1$
ii. Exponential correlogram: $\rho_d = e^{-\lambda d}$
iii. Hyperbolic correlogram: $\rho_d = tanh(d^{-3/5})$

where correlogram is the curve, or the function produced by ploting the set of correlations ρ_d for pairs of units that are d units apart against d.

Taking $L = N$ and $\lambda = 1$, the numerical values of the expected variance of the sample mean under each of SRS, RLSS, CSS, NPSS, MRSS and RLSSM are obtained, using R package for statistical computing, for the three types of correlograms as in Tables (6), (7) and(8), respectively.

Clearly, the numerical results in Tables (6), (7) and (8), show that the proposed sampling procedure is better than SRS for the three different types of correlogram regardless the number of random starts. Compared with RLSS procedure, the suggested sampling design has higher expected variance in all cases. However, the propsed design still has the merit over RLSS by handling the two main statistical issues of LSS simultaneously. On the other hand, the suggested sampling procedure is more efficient than CSS in most cases, espcially for large population and sample sizes, for populations with linear correlogram.

Table 6. The Expected variances corresponding to six sampling procedures, namely, SRS, RLSS, CSS, NPSS, MRSS and RLSSM for populations exhibt a linear correlogram ($\rho_d = 1-(d/N)$).

N	n	u,a	t_1,t_2	σ^2_{SRS}	σ^2_{RLSS}	σ^2_{CSS}	σ^2_{NPSS}	σ^2_{MRSS}	σ^2_{RLSSM}
20	8	12, 4	2, 2	0.0263	0.0050	0.0175	0.0185	0.0156	0.0087
36	8	6, 2	2, 2	0.0333	0.0051	0.0090	0.0108	0.0181	0.0094
36	8	16, 4	2, 2	0.0333	0.0051	0.0090	0.0181	0.0181	0.0094
36	10	4, 2	2, 2	0.0247	0.0033	0.0072	0.0071	0.0104	0.0058
36	10	4, 2	2, 3	0.0247	0.0033	0.0072	0.0071	0.0104	0.0079
40	12	8, 4	2, 2	0.0199	0.0023	0.0054	0.0105	0.0142	0.0040
40	12	8, 4	4, 2	0.0199	0.0023	0.0054	0.0105	0.0142	0.0055
42	12	10, 4	2, 2	0.0203	0.0023	0.0089	0.0082	0.0106	0.0040
42	12	10, 4	3, 3	0.0203	0.0023	0.0089	0.0082	0.0106	0.0058
46	10	6, 2	2, 2	0.0267	0.0033	0.0057	0.0066	0.0107	0.0060
46	10	6, 2	2, 3	0.0267	0.0033	0.0057	0.0066	0.0107	0.0080
48	14	8, 4	2, 2	0.0172	0.0017	0.0067	0.0078	0.0104	0.0030
48	14	8, 4	4, 3	0.0172	0.0017	0.0067	0.0078	0.0104	0.0047
50	14	10, 4	2, 2	0.0175	0.0030	0.0049	0.0059	0.0076	0.0030
50	14	10, 4	2, 4	0.0175	0.0030	0.0049	0.0059	0.0076	0.0050
100	22	10, 4	2, 2	0.0119	0.0007	0.0031	0.0034	0.0051	0.0012
100	30	16, 2	2, 2	0.0079	0.0007	0.0037	0.0034	0.0060	0.0006
100	40	4, 8	2, 2	0.0051	0.0002	0.0135	0.0086	0.0049	0.0003
200	60	26, 2	2, 2	0.0039	0.0001	0.0034	0.0032	0.0037	0.0002
200	80	44, 4	2, 2	0.0025	0.0001	0.0134	0.0120	0.0041	0.0001
200	80	44, 4	4, 4	0.0025	0.0001	0.0134	0.0120	0.0041	0.0002

Table 7. The Expected variances corresponding to six sampling procedures, namely, SRS, RLSS, CSS, NPSS, MRSS and RLSSM for populations exhibt an exponential correlogram ($\rho_d = e^{-d}$).

N	n	u,a	t_1,t_2	σ^2_{SRS}	σ^2_{RLSS}	σ^2_{CSS}	σ^2_{NPSS}	σ^2_{MRSS}	σ^2_{RLSSM}
20	8	12, 4	2, 2	0.0708	0.0469	0.0512	0.0583	0.0585	0.0623
36	8	6, 2	2, 2	0.0941	0.0714	0.0699	0.0783	0.0839	0.0859
36	8	16, 4	2, 2	0.0941	0.0714	0.0699	0.0839	0.0839	0.0859
36	10	4, 2	2, 2	0.0699	0.0490	0.0649	0.0540	0.0575	0.0618
36	10	4, 2	2, 3	0.0699	0.0490	0.0649	0.0540	0.0575	0.0649
40	12	8, 4	2, 2	0.0567	0.0384	0.0377	0.0469	0.0515	0.0491
40	12	8, 4	4, 2	0.0567	0.0384	0.0377	0.0469	0.0515	0.0526
42	12	10, 4	2, 2	0.0579	0.0399	0.0402	0.0466	0.0493	0.0507
42	12	10, 4	3, 3	0.0579	0.0399	0.0402	0.0466	0.0493	0.0545
46	10	6, 2	2, 2	0.0763	0.0575	0.0622	0.0618	0.0654	0.0691
46	10	6, 2	2, 3	0.0763	0.0575	0.0622	0.0618	0.0654	0.0715
48	14	8, 4	2, 2	0.0494	0.0336	0.0336	0.0397	0.0431	0.0428
48	14	8, 4	4, 3	0.0494	0.0336	0.0336	0.0397	0.0431	0.0469
50	14	10, 4	2, 2	0.0502	0.0345	0.0352	0.0394	0.0414	0.0438
50	14	10, 4	2, 4	0.0502	0.0345	0.0352	0.0394	0.0414	0.0470
100	22	10, 4	2, 2	0.0350	0.0256	0.0327	0.0281	0.0301	0.0311
100	30	16, 2	2, 2	0.0231	0.0151	0.0151	0.0153	0.0207	0.0194
100	40	4, 8	2, 2	0.0140	0.0089	0.0110	0.0154	0.0119	0.0120
200	60	26, 2	2, 2	0.0116	0.0075	0.0076	0.0076	0.0103	0.0097
200	80	44, 4	2, 2	0.0075	0.0044	0.0055	0.0054	0.0060	0.0060
200	80	44, 4	4, 4	0.0075	0.0044	0.0055	0.0054	0.0060	0.0070

Table 8. The Expected variances corresponding to six sampling procedures, namely, SRS, RLSS, CSS, NPSS, MRSS and RLSSM for populations exhibt a hyperbolic correlogram ($\rho_d = tanh(d^{-3/5})$).

N	n	u,a	t_1,t_2	σ^2_{SRS}	σ^2_{RLSS}	σ^2_{CSS}	σ^2_{NPSS}	σ^2_{MRSS}	σ^2_{RLSSM}
20	8	12, 4	2, 2	0.0468	0.0170	0.0292	0.0340	0.0318	0.0291
36	8	6, 2	2, 2	0.0689	0.0303	0.0322	0.0396	0.0494	0.0457
36	8	16, 4	2, 2	0.0689	0.0303	0.0322	0.0494	0.0494	0.0457
36	10	4, 2	2, 2	0.0512	0.0192	0.0284	0.0258	0.0309	0.0306
36	10	4, 2	2, 3	0.0512	0.0192	0.0284	0.0258	0.0309	0.0355
40	12	8, 4	2, 2	0.0421	0.0147	0.0171	0.0272	0.0337	0.0238
40	12	8, 4	4, 2	0.0421	0.0147	0.0171	0.0272	0.0337	0.0285
42	12	10, 4	2, 2	0.0433	0.0154	0.0214	0.0250	0.0291	0.0247
42	12	10, 4	3, 3	0.0433	0.0154	0.0214	0.0250	0.0291	0.0303
46	10	6, 2	2, 2	0.0579	0.0241	0.0290	0.0302	0.0365	0.0363
46	10	6, 2	2, 3	0.0579	0.0241	0.0290	0.0302	0.0365	0.0412
48	14	8, 4	2, 2	0.0377	0.0128	0.0174	0.0224	0.0272	0.0207
48	14	8, 4	4, 3	0.0377	0.0128	0.0174	0.0224	0.0272	0.0267
50	14	10, 4	2, 2	0.0385	0.0132	0.0166	0.0205	0.0236	0.0212
50	14	10, 4	2, 4	0.0385	0.0132	0.0166	0.0205	0.0236	0.0269
100	22	10, 4	2, 2	0.0291	0.0103	0.0156	0.0149	0.0187	0.0158
100	30	16, 2	2, 2	0.0192	0.0055	0.0087	0.0084	0.0152	0.0090
100	40	4, 8	2, 2	0.0123	0.0029	0.0145	0.0162	0.0090	0.0051
200	60	26, 2	2, 2	0.0102	0.0027	0.0055	0.0052	0.0082	0.0044
200	80	44, 4	2, 2	0.0066	0.0014	0.0106	0.0093	0.0053	0.0025
200	80	44, 4	4, 4	0.0066	0.0014	0.0106	0.0093	0.0053	0.0037

For exponential and hyperbolic correlograms, CSS is more efficient than the proposed procedure in most cases. Considering the NPSS and MRSS procedures, our procedure is superior to both of NPSS and MRSS for populations with linear correlogram. In case of exponential correlogram, these two procedures outperform our procedure in most cases. For populations exhibit hyperbolic correlogram the proposed sampling procedure outperforms the two procedures for two random starts from each of the two strata and it becomes less efficient for larger numbers of random starts.

It is worth noting that for small samples, when we increase the sample size while fixing the population size, we will have a gain in precision for all of the six sampling procedures. On the other hand, for large populations, only the two remainder systematic sampling designs will have gains in effeciency while the expected variance of the other designs increases.

4.4 On the Choice of the Number of Random Starts

The effect of choosing certain number of random starts from each subpopulation, t_1 and t_2, on the efficiency of the proposed sampling procedure is of considerable interest. From the numerical study introduced in the previous two sections, it can be clearly noticed that with the increase of the chosen random starts, the efficiency of the proposed sampling design decreases considerably. Also, an increase of a single random start in the number of random strats from the second subpopulation, t_2, while fixing the number of random starts from the first subpopulation, t_1, has a higher effect on the efficiency of the proposed procedure than increasing t_1 while fixing t_2.

5

Conclusions and Future Work

Remainder linear systematic sampling with multiple random starts is an extension of the remainder linear systematic sampling procedure of Chang and Huang (2000). Similar to RLSS, RLSSM can be used when the population size is not a multiple of the sample size; additionally, it provides an unbiased estimator for the variance of the sample mean. Thus, the proposed sampling procedure can handle the two main statistical issues of the usual linear systematic sampling simultaneously. In the same context, our sampling design is not only easily applicable in the practice, but it can also be considered as a general systematic sampling design that can produce many other sampling procedures as special cases. For example, LSS, RLSS and MSSS.

It is found that the proposed design outperforms all other mentioned designs, except RLSS, for populations that exhibit perfect linear trend. It has the same efficiency of RLSS and more efficient than the other designs for randomly ordered populations in some cases. For auto-correlated populations, the proposed design is more efficient than the other designs, except RLSS, in case of linear correlogram. Both of RLSS and CSS outperform the proposed design for populations with hyperbolic correlogram in most cases.

Focusing on the choice of the number of random starts from each subpopulation, the numerical study showed that increasing the number of random starts in any of the two subpopulations leads to a substantial loss of efficiency of the proposed multi-start sampling procedure.

To conclude, if the statistician has the choice between the six sampling procedures involved in the performance comparisons, it is suggested to choose the remainder linear systematic sampling with multiple random starts as the variance of the sample mean can be estimated unbiasedly regardless the form of the population and the relation between N and n. However, the remainder linear systematic sampling procedure is considerably more efficient than the multi-start version. Thus, if the statistician can find at least a consistent estimator for the sampling variance, it might be worth to use the single start remainder linear systematic sampling procedure.

From this study, two main points can be considered for future research. First, by fol-

lowing the same rationale of this study, the sampling variance of the CSS design can be estimated unbiasedly through incorporating the idea of multiple random starts into this design. Second, motivated by Sampath (2012), the finite population variance of the study variable, Y, can be estimated under the proposed RLSSM design by utilizing the idea of multiple random starts.

Bibliography

Bellhouse, D. R. and Rao, J. N. K. (1975). Systematic sampling in the presence of a trend. *Biometrika* **62**, 694-697.

Chandra, K. S., Sampath, S. and Balasubramani, G. K. (1991). Markov sampling for finite populations. *Biomertika* **79**, 210-213.

Chang, H. and Huang, K. (2000). Remainder linear systematic sampling. *Sankhya* **B. 62**, 249-256.

Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling: Theory and Methods,* 2nd ed. London: Chapman & Hall/CRC.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics* **17**, 164-177.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

Food and Agriculture organization of the United Nations (2010). Global Forest Resources Assessment 2010: Main Report. Rome **163**.

Gautschi, W. (1957). Some remarks on systematic sampling. *Annals of Mathematical Statistics* **28**, 385-394.

Hadi, A. S. (1996). *Matrix Algebra as a Tool*. Belmont, California: Duxbury Press.

Horvitz, D.G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

Huang, K. (2004). Mixed random systematic sampling designs. *Metrika* **59**, 1-11.

Kao, F., Leu, C. and Ko, C. (2011). Remainder Markov systematic sampling. *Journal of Statistical Planning and Inference* **141**, 3595- 3604.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Lahiri, D. B. (1951). A method for selection providing unbiased estimates. *International Statistical Association Bulletin* **33**, 133-140.

Leu, C. and Tsui, K. (1996). New partially systematic sampling. *Statistica Sinica* **6**, 617-630.

Levy S. and Lemeshow S. (2008). *Sampling of populations: Methods and Applications*. New York: John Wiley & Sons.

Madow, W.G. (1953). On the theory of systematic sampling, III. Comparison of centered and random start systematic sampling. *Annals of Mathematical Statistics* **24**, 101-106.

Madow, W.G. and Madow, L. H. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics* **25**, 1-24.

Montanari, G. E. and Bartolucci, F. (1998).On estimating the variance of the systematic sample mean. *Journal of the Italian Statistical Society* **7**, 185-196.

Opsomer, J. D., Francisco-Fernandez, M. and Li, X. (2012). Model-based nonparametric variance estimation for systematic sampling. *Scandinavian Journal of Statistics* **39**, 538-542

Sampath, S. and Ammani, S. (2010). Some systematic sampling strategies using multiple random starts. *Pakistan Journal of Statistics and Operations Research* **6**, 149-162.

Sampath, S. and Ammani, S. (2012): Finite-population variance estimation under systematic sampling schemes with multiple random starts. *Journal of Statistical Computation and Simulation* **82**, 1207-1221.

Sampath, S. and Uthayakumaran, N. (1998). Markov systematic sampling. *Biometrical Journal* **40**, 883-895.

Sengupta, S. and Chattopadhyay, S. (1987). A note on circular systematic sampling. *Sankhya* **B.46**, 186-187.

Sethi, V. K. (1965). On optimum pairing of units. *Sankhya* **B. 27**, 315-320.

Singh, D., Jindal, K. and Garg, J. N.(1968). On modified systematic sampling. *Biometrika* **55**, 541-546.

Singh, D. and Singh, P. (1977). New systematic sampling. *Journal of Statistical Planning and Inference* **1**, 163-177.

Sudakar, K. (1978). A note on circular systematic sampling design. *Sankhya* **40**, 72-73.

Wolter, K. M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* **79**, 781-790.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Yates, F. (1948). Systematic sampling. *Philosophical Transactions of the Royal Society of London* **241**, 345-377.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* **15**, 253-261.

Zinger, A. (1964). Systematic sampling in forestry. *Biometrics* **20**, 553-565.

Zinger, A. (1980). Variance estimation in partially systematic sampling. *Journal of the American Statistical Association* **75**, 206-211.

APPENDIX

R- programs for the numerical study

1. Computing the value of sigma-squared (CSS) when $N = 20, n = 8$ for populations with linear trend:

$N < -20$
$n < -8$
$k < -2$
$A < -0$
$B < -0$
$C < -0$
$sigma2_{CSS} < -0$
$for(i\ in\ 1:N)\{A < -A + ((i - ((N+1)/2))^2)$
$\}$
$for(i\ in\ 1:(N-1))\{$
$for(j\ in\ (i+1):N)\{B < -B + ((i - ((N+1)/2)) * (j - ((N+1)/2)))$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:(n-2))\{$
$for(j\ in\ (i+1):(n-1))\{$
$C < -if((t + (i*k)) <= N \& (t + (j*k)) <= N)\{C + ((t + (i*k) - ((N+1)/2)) * (t + (j*k) - ((N+1)/2)))$
$\}\ else\ if((t + (i*k)) <= N \& (t + (j*k)) > N)\{C + ((t + (i*k) - ((N+1)/2)) * (t + (j*k) - N - ((N+1)/2)))$
$\}\ else\ if((t + (i*k)) > N \& (t + (j*k)) <= N)\{C + ((t + (i*k) - N - ((N+1)/2)) * (t + (j*k) - ((N+1)/2)))$
$\}\ else\{C + ((t + (i*k) - N - ((N+1)/2)) * (t + (j*k) - N - ((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$
$(1/N) * ((1/n) - (1/N)) * A + (2/(N * (n^2))) * C - (2/(N^2)) * B$

2. Computing the value of sigma-squared (CSS) when $N = 20, n = 8$ for populations with exponential correlogram:

$N < -20$
$n < -8$
$k < -2$
$B < -0$
$D < -0$
$sigma2_{CSS} < -0$
$for(d\ in\ 1:(N-1))\{rhod < -exp(-d)$
$B < -B + (N-d) * rhod$
$\}$
$for(t\ in\ 1:N)\{$

$for(i\ in\ 0:(n-2))\{$
$for(j\ in\ (i+1):(n-1))\{$
$D<-if(((t+i*k)<=N)\&((t+j*k)<=N))\{D+exp(-(abs((t+i*k)-(t+j*k))))$
$\}else\ if(((t+i*k)<=N)\&((t+j*k)>N))\{D+exp(-(abs((t+i*k)-(t+j*k-N))))$
$\}else\ if(((t+i*k)>N)\&((t+j*k)<=N))\{D+exp(-(abs((t+i*k-N)-(t+j*k))))$
$\}else\ \{D+exp(-(abs((t+i*k-N)-(t+j*k-N))))$
$\}$
$\}$
$\}$
$\}$
$sigma2_{CSS}<-((1/n)-(1/N))+(2/(N*(n^2)))*D-(2/(N^2))*B$

3. Computing the value of sigma-squared (MRSS) when $N=20, n=8$ for populations with linear trend:

$N<-20$
$n<-7$
$k<-2$
$r<-6$
$A<-0$
$B<-0$
$C<-0$
$D<-0$
$E<-0$
$for(i\ in\ 1:N)\{A<-A+((i-((N+1)/2))^2)$
$\}$
$for(i\ in\ 1:(N-1))\{$
$for(j\ in\ (i+1):N)\{B<-B+((i-((N+1)/2))*(j-((N+1)/2)))$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:((n-r)*k-2))\{$
$for(j\ in\ (i+1):((n-r)*k-1))\{$
$C<-if((t+i)<=N\&(t+j)<=N)\{C+((t+i-((N+1)/2))*(t+j-((N+1)/2)))$
$\}elseif((t+i)<=N\&(t+j)>N)\{C+((t+i-((N+1)/2))*(t+j-N-((N+1)/2)))$
$\}elseif((t+i)>N\&(t+j)<=N)\{C+((t+i-N-((N+1)/2))*(t+j-((N+1)/2)))$
$\}else\{C+((t+i-N-((N+1)/2))*(t+j-N-((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:((n-r)*k-1))\{$
$for(j\ in\ 1:(r))\{$

$D < -if((i+t) <= N\&(t+j*(k+1)+(n-r)*k-1) <= N)\{D+(((i+t)-((N+1)/2))*(t+j*(k+1)+(n-r)*k-1-((N+1)/2)))$
$\}elseif((i+t) <= N\&(t+j*(k+1)+(n-r)*k-1) > N)\{D+(((i+t)-((N+1)/2))*(t+j*(k+1)+(n-r)*k-1-N-((N+1)/2)))$
$\}elseif((i+t) > N\&(t+j*(k+1)+(n-r)*k-1) <= N)\{D+(((i+t)-N-((N+1)/2))*(t+j*(k+1)+(n-r)*k-1-((N+1)/2)))$
$\}else\{D+(((i+t)-N-((N+1)/2))*((t+j*(k+1)+(n-r)*k-1)-N-((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 1:(r-1))\{$
$for(j\ in\ (i+1):r)\{$
$E < -if((t+i*(k+1)+(n-r)*k-1) <= N\&(t+j*(k+1)+(n-r)*k-1) <= N)\{E+((t+i*(k+1)+(n-r)*k-1-((N+1)/2))*((t+j*(k+1)+(n-r)*k-1)-((N+1)/2)))$
$\}elseif((t+i*(k+1)+(n-r)*k-1) <= N\&(t+j*(k+1)+(n-r)*k-1) > N)\{E+(((t+i*(k+1)+(n-r)*k-1)-((N+1)/2))*((t+j*(k+1)+(n-r)*k-1)-N-((N+1)/2)))$
$\}elseif((t+i*(k+1)+(n-r)*k-1) > N\&(t+j*(k+1)+(n-r)*k-1) <= N)\{E+(((t+i*(k+1)+(n-r)*k-1)-N-((N+1)/2))*((t+j*(k+1)+(n-r)*k-1)-((N+1)/2)))$
$\}else\{E+(((t+i*(k+1)+(n-r)*k-1)-N-((N+1)/2))*((t+j*(k+1)+(n-r)*k-1)-N-((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$
$(1/N)*((1/n)-(1/N))*A+((n-r-1)/(k*((n-r)*k-1)))*(2/(N*(n^2)))*(C)+(1/k)*(2/(N*(n^2)))*D+(2/(N*(n^2)))*E-(2/(N^2))*B$

4. Computing the value of sigma-squared (MRSS) when $N = 20, n = 8$ for populations with hyperbolic correlogram :

$N < -20$
$n < -8$
$k < -2$
$r < -4$
$A < -0$
$B < -0$
$C < -0$
$D < -0$
$for(d\ in\ 1:(N-1))\{rhod < -tanh(d^(-3/5))$
$B < -B+(N-d)*rhod$
$\}$

$$
\begin{aligned}
&for(t\ in\ 1:N)\{\\
&for(i\ in\ 1:(r-1))\{\\
&for(j\ in\ (i+1):r)\{\\
&D<-if(((t+i*(k+1)+(n-r)*k-1)<=N)\&((t+j*(k+1)+(n-r)*k-1)<=\\
&N))\{D+tanh((abs((t+i*(k+1)+(n-r)*k-1)-(t+j*(k+1)+(n-r)*k-1)))^(-\\
&3/5))\\
&\}elseif(((t+i*(k+1)+(n-r)*k-1)<=N)\&((t+j*(k+1)+(n-r)*k-1)>\\
&N))\{D+tanh((abs((t+i*(k+1)+(n-r)*k-1)-(t+j*(k+1)+(n-r)*k-1-\\
&N)))^(-3/5))\\
&\}elseif(((t+i*(k+1)+(n-r)*k-1)>N)\&((t+j*(k+1)+(n-r)*k-1)<=\\
&N))\{D+tanh((abs((t+i*(k+1)+(n-r)*k-1-N)-(t+j*(k+1)+(n-r)*k-\\
&1)))^(-3/5))\\
&\}else\{D+tanh((abs((t+i*(k+1)+(n-r)*k-1-N)-(t+j*(k+1)+(n-r)*\\
&k-1-N)))^(-3/5))\\
&\}\\
&\}\\
&\}\\
&\}\\
&for(t\ in\ 1:N)\{\\
&for(i\ in\ 0:((n-r)*k-1))\{\\
&for(j\ in\ 1:r)\{\\
&C<-if(((t+i)<=N)\&((t+j*(k+1)+(n-r)*k-1)<=N))\{C+tanh((abs((t+\\
&i)-(t+j*(k+1)+(n-r)*k-1)))^(-3/5))\\
&\}elseif(((t+i)<=N)\&((t+j*(k+1)+(n-r)*k-1)>N))\{C+tanh((abs((t+i)-\\
&(t+j*(k+1)+(n-r)*k-1-N)))^(-3/5))\\
&\}elseif(((t+i)>N)\&((t+j*(k+1)+(n-r)*k-1)<=N))\{C+tanh((abs((t+i-\\
&N)-(t+j*(k+1)+(n-r)*k-1)))^(-3/5))\\
&\}else\{C+tanh((abs((t+i-N)-(t+j*(k+1)+(n-r)*k-1-N)))^(-3/5))\\
&\}\\
&\}\\
&\}\\
&\}\\
&for(t\ in\ 1:N)\{\\
&for(i\ in\ 0:((n-r)*k-2))\{\\
&for(j\ in\ (i+1):((n-r)*k-1))\{\\
&A<-if(((t+i)<=N)\&((t+j)<=N))\{A+tanh((abs((t+j)-(t+i)))^(-3/5))\\
&\}elseif(((t+i)<=N)\&((t+j)>N))\{A+tanh((abs((t+j-N)-(t+i)))^(-3/5))\\
&\}elseif(((t+i)>N)\&((t+j)<=N))\{A+tanh((abs((t+j)-(t+i-N)))^(-3/5))\\
&\}else\{A+tanh((abs((t+j-N)-(t+i-N)))^(-3/5))\\
&\}\\
&\}\\
&\}\\
&\}\\
&((1/n)-(1/N))+(2/(N*(n^2)))*(((n-r-1)/(k*((n-r)*k-1)))*A+(1/k)*C+\\
&D)-(2/(N^2))*B
\end{aligned}
$$

5. Computing the value of sigma-squared (NPSS) when $N = 20, n = 8$ for populations with linear trend :

$N < -20$
$n < -8$
$k < -2$
$u < -12$
$a < -4$
$A < -0$
$B < -0$
$C < -0$
$D < -0$
$E < -0$
$for(i\ in\ 1:N)\{A < -A + ((i - ((N+1)/2))^2)$
$\}$
$for(i\ in\ 1:(N-1))\{$
$for(j\ in\ (i+1):N)\{B < -B + ((i - ((N+1)/2)) * (j - ((N+1)/2)))$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:(u-2))\{$
$for(j\ in\ (i+1):(u-1))\{$
$C < -if((t+i) <= N \& (t+j) <= N)\{C + ((t+i-((N+1)/2)) * (t+j-((N+1)/2)))$
$\}elseif((t+i) <= N \& (t+j) > N)\{C + ((t+i-((N+1)/2)) * (t+j-N-((N+1)/2)))$
$\}elseif((t+i) > N \& (t+j) <= N)\{C + ((t+i-N-((N+1)/2)) * (t+j-((N+1)/2)))$
$\}else\{C + ((t+i-N-((N+1)/2)) * (t+j-N-((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$

for (t in 1:N) for (i in 0:(u-1)) for (j in 1:(n-a)) $D < -if((i+t) <= N \& (t + (j*k) + u - 1) <= N)\{D + (((i+t) - ((N+1)/2)) * (t + (j*k) + u - 1 - ((N+1)/2)))$
$\}elseif((i+t) <= N \& (t + (j*k) + u - 1) > N)\{D + (((i+t) - ((N+1)/2)) * (t + (j*k) + u - 1 - N - ((N+1)/2)))$
$\}elseif((i+t) > N \& (t + (j*k) + u - 1) <= N)\{D + (((i+t) - N - ((N+1)/2)) * (t + (j*k) + u - 1 - ((N+1)/2)))$
$\}else\{D + (((i+t) - N - ((N+1)/2)) * ((t + (j*k) + u - 1) - N - ((N+1)/2)))$
$\}$
$\}$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 1:(n-a-1))\{$
$for(j\ in\ (i+1):(n-a))\{$
$E < -if((t + (i*k) + u - 1) <= N \& (t + (j*k) + u - 1) <= N)\{E + ((t + (i*k) + u -$

$1-((N+1)/2))*(t+(j*k)+u-1-((N+1)/2)))$
$\}elseif((t+(i*k)+u-1)<=N\&(t+(j*k)+u-1)>N)\{E+(((t+(i*k)+u-1)-$
$((N+1)/2))*((t+(j*k)+u-1)-N-((N+1)/2)))$
$\}elseif((t+(i*k)+u-1)>N\&(t+(j*k)+u-1)<=N)\{E+(((t+(i*k)+u-1)-$
$N-((N+1)/2))*((t+(j*k)+u-1)-((N+1)/2)))$
$\}else\{E+(((t+(i*k)+u-1)-N-((N+1)/2))*((t+(j*k)+u-1)-N-((N+$
$1)/2)))$
$\}$
$\}$
$\}$
$\}$
$(1/N)*((1/n)-(1/N))*A+((a*(a-1))/(u*(u-1)))*(2/(N*(n^2)))*(C)+(a/u)*$
$(2/(N*(n^2)))*D+(2/(N*(n^2)))*E-(2/(N^2))*B$

6. Computing the value of sigma-squared (NPSS) when $N = 20, n = 8$ for populations with linear correlogram :

$N<-20$
$n<-8$
$u<-12$
$a<-4$
$k<-2$
$L<-N$
$A<-0$
$B<-0$
$C<-0$
$D<-0$
$for(d\ in\ 1:(N-1))rhod<-(1-(d/L))$
$B<-B+(N-d)*rhod$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 1:(n-a-1))\{$
$for(j\ in\ (i+1):(n-a))\{$
$D<-if(((t+i*k+u-1)<=N)\&((t+j*k+u-1)<=N))\{D+1-((abs((t+i*$
$k+u-1)-(t+j*k+u-1)))/L)$
$\}elseif(((t+i*k+u-1)<=N)\&((t+j*k+u-1)>N))\{D+1-((abs((t+i*k+$
$u-1)-(t+j*k+u-1-N)))/L)$
$\}elseif(((t+i*k+u-1)>N)\&((t+j*k+u-1)<=N))\{D+1-((abs((t+i*k+$
$u-1-N)-(t+j*k+u-1)))/L)$
$\}else\{D+1-((abs((t+i*k+u-1-N)-(t+j*k+u-1-N)))/L)$
$\}$
$\}$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:(u-1))\{$
$for(j\ in\ 1:(n-a))\{$
$C<-if(((t+i)<=N)\&((t+j*k+u-1)<=N))\{C+1-((abs((t+i)-(t+j*k+$

$u-1)))/L)$
$\}elseif(((t+i)<=N)\&((t+j*k+u-1)>N))\{C+1-((abs((t+i)-(t+j*k+u-1-N)))/L)$
$\}elseif(((t+i)>N)\&((t+j*k+u-1)<=N))\{C+1-((abs((t+i-N)-(t+j*k+u-1)))/L)$
$\}else\{C+1-((abs((t+i-N)-(t+j*k+u-1-N)))/L)\ \}$
$\}$
$\}$
$\}$
$for(t\ in\ 1:N)\{$
$for(i\ in\ 0:(u-2))\{$
$for(j\ in\ (i+1):(u-1))\{$
$A<-if(((t+j)<=N)\&((t+i)<=N))\{A+1-((abs((t+j)-(t+i)))/L)$
$\}elseif(((t+i)<=N)\&((t+j)>N))\{A+1-((abs((t+j-N)-(t+i)))/L)$
$\}elseif(((t+i)>N)\&((t+j)<=N))\{A+1-((abs((t+j)-(t+i-N)))/L)$
$\}else\{A+1-((abs((t+j-N)-(t+i-N)))/L)$
$\}$
$\}$
$\}$
$\}$
((1/n)-(1/N))+(2/(N*(n^2)))$*(((a*(a-1))/(u*(u-1)))*A+(a/u)*C+D)-(2/(N^2))*B$

7. Computing the value of sigma-squared (RLSSM) when $N=20, n=8, t_1=2, t_2=2$ for populations with exponential correlogram :

$N<-20$
$n<-8$
$k<-2$
$r<-4$
$t1<-2$
$t2<-2$
$A<-0$
$B<-0$
$C<-0$
$D<-0$
$sigmaRLSSM<-0$
$for(d\ in\ 1:(((n-r)*k)-1))\{rhod<-exp(-d)$
$A<-A+((2*(k-1))/(t1*k-1))*(((n-r)*k-d)*rhod)$
$\}$
$for(d1\ in\ 1:(((n-r)/t1)-1))\{rhodt1k<-exp(-(d1*t1*k))$
$B<-B+(2*(t1^2)*(k^2)*(k-1)/(t1*k-1))*((((n-r)/t1)-d1)*rhodt1k)$
$\}$
$for(d2\ in\ 1:(r*(k+1)-1))\{rhod<-exp(-(d2))$
$C<-C+(2*k/(t2*(k+1)-1))*((r*(k+1)-d2)*rhod)$
$\}$
$for(d3\ in\ 1:((r/t2)-1))\{rhodt2(k+1)<-exp(-(d3*t2*(k+1)))$
$D<-D+(2*t2^2*k*((k+1)^2)/(t2*(k+1)-1))*(((r/t2)-d3)*rhodt2(k+1))$

}

$$sigmaRLSSM < -sigmaRLSSM + ((1/(N^2)) * (k * (N - n + r) - A + B - C + D))$$

VITA

SAYED A. MOSTAFA ABDELMEGEED

Candidate for the Degree of

Master of Science

Thesis: REMAINDER LINEAR SYSTEMATIC SAMPLING WITH MULTI-

PLE RANDOM STARTS

Major Field: Statistics

Biographical:

Education:
Completed the requirements for the Master of Science in Statistics at Oklahoma State University, Stillwater, Oklahoma in May, 2014.

Completed the requirements for the Bachelor of Science in Statistics at Cairo University, Cairo, Egypt in 2010.

Experience:
July, 2009 - September 2009: Training in the Information and Decision Support Center, Cabinet, Egypt.

September, 2010 - August, 2012: Teaching Assistant in Department of Statistics at Cairo University.

February, 2010 - May, 2010: Working in CAPMAS, Egypt as a member in a statistical consulting team.

August, 2012 - December, 2012: Teaching Assistant in Department of Statistics at Oklahoma State University.

December, 2012 - Now: Instructor in Department of Statistics at Oklahoma State University.

Professional Memberships:

March 2013 - Now: Founding Member and selected for Lifetime Membership in the Delta Epsilon Iota Academic Honor Society.