

परियोजना रिपोर्ट PROJECT REPORT

द्वि-स्तरीय प्रतिचयन अभिकल्पना के अंतर्गत डोमेन कैलिब्रेशन
आकलकों पर अध्ययन

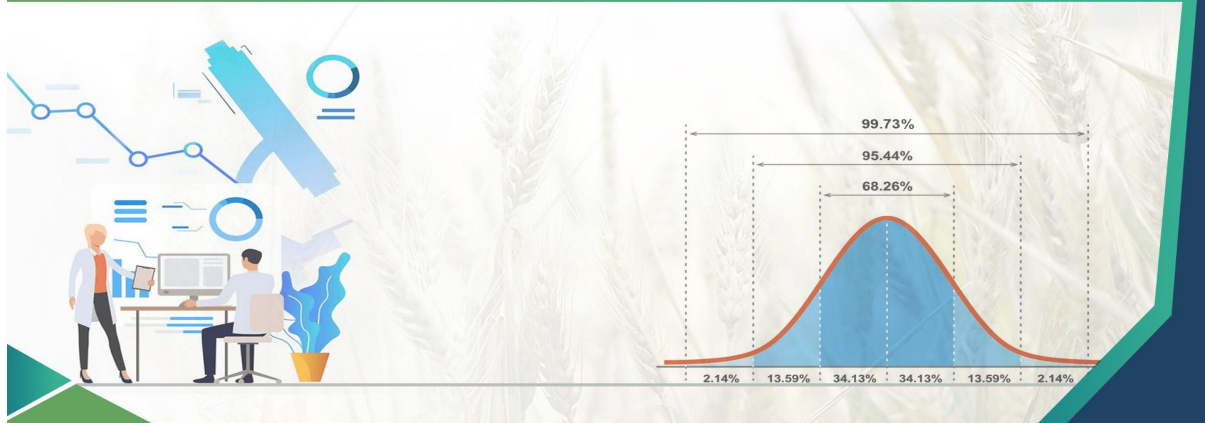
A Study on Domain Calibration Estimators under Two Stage Sampling Design



भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली – 110012
ICAR-Indian Agricultural Statistics Research Institute Library Avenue,
Pusa, New Delhi - 110012



A Pioneer Institute of ICAR undertaking Research, Teaching and Training in Agricultural Statistics, Computer Application and Bioinformatics



कौस्तव आदित्य
राजू कुमार
पंकज दास

Kaustav Aditya
Raju Kumar
Pankaj Das



प्रतिदर्श सर्वेक्षण प्रभाग
Division of Sample Surveys
भा. कृ. अनु. प.-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
ICAR-Indian Agricultural Statistics Research Institute
लाइब्रेरी एवेन्यू, पूसा नई दिल्ली-110012
Library Avenue, Pusa, New Delhi – 110012
<https://iasri.icar.gov.in>
2023



आमुख

प्रतिदर्श सर्वेक्षणों में, परिमित समष्टि में सहायक चर की जानकारी का उपयोग परिमित समष्टि कुल या औसत या वितरण के आकलकों की परिशुद्धता बढ़ाने के लिए किया जाता है। आमतौर पर, समष्टि कुल या औसत के आकलकों की परिशुद्धता बढ़ाने के लिए, अनुपात एवं रिग्रेशन आकलक सहायक चर की जानकारी का उपयोग करते हैं। देविल और सरंडल (1992) द्वारा प्रस्तावित कैलिब्रेशन प्रक्रिया एक ऐसी तकनीक है जिसमें सर्वेक्षण अनुमान के लिए सहायक चर की जानकारी को कुशलतापूर्वक उपयोग किया जाता है, विशेष रूप से जब जनसंख्या स्तर पर सहायक चर की जानकारी उपलब्ध हो। कभी कभी यह भी देखा जाता है कि समष्टि की उपसमुच्चय या डोमेन के लिए भी अनुमान आवश्यक होते हैं। कैलिब्रेशन तकनीक से संबंधित अधिकांश शोध कार्य एक चरण या द्वी-प्रावस्था प्रतिचयन डोमेन प्राचल आकलन के लिए ही सीमित है। लेकिन बड़े पैमाने के सर्वेक्षणों में द्वी-चरण या बहुचरण प्रतिचयन का उपयोग किया जाता है। इसलिए द्वी-चरण प्रतिचयन अभिकल्पना के अंतर्गत डोमेन कैलिब्रेशन आकलकों को विकसित करने की आवश्यकता थी। इस शोध परियोजना का शीर्षक था "द्वि-स्तरीय प्रतिचयन अभिकल्पना के अंतर्गत डोमेन कैलिब्रेशन आकलकों पर अध्ययन", जिसका उद्देश्य जटिल सहायक चर जानकारी की उपस्थिति में द्वी-चरण प्रतिचयन डिज़ाइन के अंतर्गत डोमेन कैलिब्रेशन आकलकों को विकसित करना था। यह शोध कार्य भा.कृ.अनु.प.-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली की अनुसंधान परियोजना के अंतर्गत किया गया और द्वि-स्तरीय प्रतिचयन अभिकल्पना के अंतर्गत डोमेन कैलिब्रेशन आकलक विकसित किये गए।

कौस्तव आदित्य

राजू कुमार

पंकज दास

PREFACE

In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of finite population total or mean or distribution function. In the simplest settings, ratio and regression estimators incorporate known finite population parameters of auxiliary variables. The Calibration Approach proposed by Deville and Sarndal (1992) is one of the other techniques widely used for making efficient use of auxiliary information in survey estimation when there was availability of population level auxiliary information. Sometimes estimates were also needed for various sub-populations or domains within the populations. Most of the works related to calibration approach was mostly restricted for estimation of the domain parameters in Single Stage or Two Phase sampling designs. But in case of large scale surveys two stage or multistage sampling designs were used. Hence there was a need to develop domain calibration estimators for two stage sampling design. This research project entitled “A study on domain calibration estimation under two stage sampling design” was undertaken at the Indian Agricultural Statistics Research Institute, New Delhi with objective to develop the domain calibration estimators under two stage sampling design in the presence of complex auxiliary information.

Kaustav Aditya

Raju Kumar

Pankaj Das

आभार

लेखकगण, भा.कृ.अनु.प.-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान के निदेशक डॉ. राजेंद्र प्रसाद के प्रति अपनी गहरी कृतज्ञता व्यक्त करते हैं कि उन्होंने इस अध्ययन को संचालित करने में सहायता और मूल्यवान मार्गदर्शन के साथ-साथ आवश्यक सुविधाएँ प्रदान की।

लेखकगण, प्रतिदर्श सर्वेक्षण प्रभाग के अध्यक्ष डॉ. तौक़ीर अहमद का परियोजना में अपने अमूल्य सुझावों और विचारों के माध्यम से योगदान करने पर आभार व्यक्त करते हैं।

लेखकगण, इस शोध के कार्य में मूल्यवान मार्गदर्शन और सहायता प्रदान करने वाले प्रभाग के सभी वैज्ञानिकों और कर्मचारियों का आभार व्यक्त करना चाहते हैं।

लेखकगण

ACKNOWLEDGEMENTS

The Authors express their deep sense of gratitude to Dr. Rajender Parsad, Director, ICAR-Indian Agricultural Statistics Research Institute (ICAR-LASRI), New Delhi for extending support and valuable guidance as well as necessary facilities in carrying out this study.

The Authors would also like to place on record the cooperation Dr. Tauqueer Ahmed, Head, Division of Sample Survey for giving his valuable advice in the research carried out in this project.

The Authors acknowledge the assistance provided by all other staffs of the division for their valuable guidance and assistance.

AUTHORS

CONTENTS

CHAPTER	TITLE	PAGE
1.	INTRODUCTION	1.1-1.20
1.1	Introduction	1.1-1.18
1.2	Motivation and Objective of the Research Project	1.19-1.20
1.3	Structure of the Project Report	1.20
2.	CALIBRATION ESTIMATION TECHNIQUE	2.1-2.16
2.1	Introduction	2.1-2.9
2.2	Proposed Domain Calibration Estimators under Two Stage Sampling Design	2.10-2.16
3.	EMPIRICAL STUDY	3.1-3.15
3.1	Empirical evaluation	3.1-3.3
3.2	Performance Measures	3.4-3.5
3.3	Results	3.5-3.8
3.4	Discussion	3.8-3.13
3.4	Summary of the Major Findings	3.13-3.15
4.	CONCLUSIONS AND FUTURE RESEARCH	4.1-4.4
4.1	Introduction	4.1
4.2	Major Findings	4.1-4.3
4.3	Further Research Areas	4.4
	सारांश	
	EXECUTIVE SUMMARY	
	REFERENCES	

CHAPTER 1

1. Introduction

Researchers typically use sample survey methodology to get information about the population or large aggregates by choosing and measuring a sample from that population. Because of the variability of characteristics among things within the population, researchers apply scientific sample styles within the sample choice method to scale back the chance of a distorted read of the population and that they build inferences regarding the population supported by the data from the sample survey data. That is, a survey plays a significant role in collecting information from the population. The target of sample surveys is to create inferences about a population from information present in the sample which is selected from that population. The inference may take the form of estimating a population mean (such as the mean yield of the crop) or proportion (such as the proportion of people suffering from the disease). Every observation, or item, taken from the population contains a precise quantity of data regarding the population parameter or parameters of interest. Because information costs money, the experimenter must determine how much information he or she should need. Insufficient information prevents the experimenter from creating smart estimates, whereas an excessive amount of information ends up in a waste of cash. The amount of information obtained within the sample depends on the number of things sampled and on the quantity of variation within the data. This latter issue is often controlled somewhat by the tactic of choosing the sample, known as the design of the sample survey. The design of the sample survey and also the sample size determines the amount of data within the sample pertinent to a population parameter, only if correct measurements are obtained on every sampled part.

The demand for statistical information appears to be limitless in contemporary society. Specifically, data is consistently gathered to fulfill the need for information about specific groups of elements known as finite populations. One of the primary methods used to collect such data is through sample surveys, which involve conducting a partial investigation of the finite population. In this context, the term "population" refers to a group of units defined based on the survey's objectives. The desired information about the population typically includes the total number of units, aggregate values of different characteristics, averages of various attributes, and so on. Conducting a sample survey is more cost-effective and less time-consuming compared to a complete enumeration, and it can even yield more accurate results. When referring to a set of units or a subset of the total material chosen to be representative of the entire aggregate, we use the term "sample." If the selection of the sample is governed by ascertainable laws of chance, it is referred to as a random or probability sample. In other words, a random or probability sample is drawn in a way that each unit in the population has a predetermined probability of being selected. The field of sampling theory addresses the scientific and objective procedures for selecting an appropriate sampling design, which aims to obtain a representative sample of the population as a whole. Additionally, it provides suitable estimation techniques for estimating population parameters. Sometimes, the primary objective of a sampling design is to achieve a specified level of precision while minimizing costs or to maximize precision given a fixed cost. An essential requirement for conducting a reliable survey is to provide a measure of precision for each estimate derived from the survey data.

Sampling techniques find application in surveys conducted worldwide. The primary objective of many surveys is to obtain descriptive measures pertaining to the characteristics of the entire population being studied. Such surveys are highly prevalent and crucial for generating data

necessary for national planning and socio-economic development. For instance, in the field of agriculture, data concerning crop production, land utilization, and water resources are indispensable for planning purposes. Sampling methods also play a role in various censuses. Apart from collecting certain fundamental information about every individual or area, data on different aspects are gathered through sampling. Sampling methods offer cross-checks and expedite the process of tabulation and publication of results. In business and industry, sampling techniques are extensively employed to enhance operational efficiency. They hold significance in addressing market research challenges like estimating readership numbers for news magazines and newspapers or gauging consumer responses to recently introduced products. Prominent references in this field include Yates (1953), Hansen *et al.* (1953), Kish (1969), and Cochran (1977).

1.1 Two Stage Sampling

Typically, sampling designs operate under the assumption that direct element sampling is feasible, meaning there is a sampling frame available to describe the target population and use it for sample selection. However, in many medium to large-scale surveys, this may not be the case, or obtaining a sampling frame could be prohibitively expensive. Additionally, if the population is geographically dispersed, it can result in high travel expenses for interviewers and pose challenges for effective fieldwork supervision, leading to increased non-response rates and measurement errors.

To address these issues, various sampling designs have been developed, such as cluster sampling and multistage sampling. In cluster sampling, the finite population is divided into subpopulations called clusters, and all elements within the selected clusters are enumerated. It's important to note that the efficiency of cluster sampling decreases as the cluster size increases. To improve

precision in such situations, a two-stage sampling approach is often employed. This involves first selecting clusters and then choosing a specific number of elements from each selected cluster. This process of selecting elements in the sample is known as two-stage or sub-sampling. The clusters selected at the first stage are referred to as first-stage units (fsu) or primary-stage units (psu), while the elements within clusters are called second-stage units (ssu). For example, in the case of a crop survey, fields can be considered as first-stage units, and plots within fields would be the second-stage units. The two-stage sampling procedure can be extended to three or more stages, known as multi-stage sampling, which is commonly used in large-scale surveys. Cochran (1977), Hansen *et al.* (1953), Sukhatme (1984) have discussed the application of this procedure in agricultural and population surveys.

1.2 Domain Estimation

Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario, i.e. In India where Land records/khasra registers are available. Now total number of villages (clusters) in each Tehsil (stratum) is known but the total number of villages under the constituent crop (Rice, wheat etc.) in the mixture i.e. number of villages having the crop as Pure Stand, mixture-1, mixture-2... may not be available. Further, the number of selected villages within each tehsil is fixed, but the number of selected villages within

each stratum under the crop as pure stand, mixture-1, mixture-2...is a random quantity. These different categories pure stand, mixture-1, mixture-2 ... may be considered as Domains. Domain estimation is a crucial aspect of sample surveys that allows researchers to make accurate inferences about specific subgroups or domains within a population. In many cases, the primary goal of a survey is not only to estimate population parameters but also to provide reliable estimates for smaller groups or domains of interest. These domains could be defined based on demographic characteristics, geographical regions, or any other relevant criteria. Domain estimation involves the application of statistical techniques to estimate parameters specific to these subgroups. It allows researchers to gain insights into the variations and characteristics unique to each domain, enabling more targeted and informed decision-making. By focusing on domains, survey results can be customized and tailored to address the specific needs and requirements of different population segments. Small area estimation is a special case of domain estimation. Small area estimation is a specialized technique that complements domain estimation in sample surveys. While domain estimation focuses on obtaining accurate estimates for specific subgroups or domains within a population, small area estimation takes it a step further by providing reliable estimates for smaller geographic areas or sub-regions. It addresses the challenge of limited sample sizes within these areas, which can lead to high sampling errors and less precise estimates. Small area estimation leverages both survey data and auxiliary information, such as administrative records or satellite imagery, to improve the precision of estimates at the small area level. By borrowing strength from the larger sample and the available auxiliary information, small area estimation allows researchers to obtain more robust and accurate estimates for areas with limited sample representation. This technique is particularly valuable in policy-making, resource allocation, and decision-making processes that require

granular information about specific geographic regions or subpopulations within domains. Thus, small area estimation complements domain estimation by extending the scope of precision and enabling more targeted and localized insights for decision-makers.

1.3 Auxiliary information

Before the survey's planning and execution, data on specific variables " x " are frequently accessible at the population level. Auxiliary information is the term used frequently to describe this. In real-world circumstances, there are typically two scenarios:

(i) All survey values $\{x_1, x_2, \dots, x_N\}$ are known for the entire survey population. This is the so-called comprehensive (or) complete auxiliary information.

(ii) The population totals of x or the population means x are known. Auxiliary information may be obtained from different data sources and in different forms such as census, population-based survey reports, results of previous experiments, etc. Administrative information like tax returns, business registers, and medical records may contain it. It can also be derived from earlier surveys with high sample sizes and estimations from survey data that are thought to be extremely reliable and can be used to make important population-related decisions. Preliminary analysis of aerial photographs or satellite photos might yield helpful supplemental data for natural resource inventory assessments.

In general, auxiliary data can be utilized during the estimating stage or the survey design stage, or both. For stratified sampling designs, stratum membership variables are utilized, and PPS sampling requires the variable indicating the size measure of units. The kind of estimating procedures to be employed is frequently determined by the quantity of auxiliary information that is accessible. Values of the response variable y and the auxiliary variable x were automatically gathered for units included in the sample if the auxiliary variable x was present. In this situation,

survey data consist of both y and x plus the auxiliary population information on x . Using auxiliary information mostly serves to improve estimation precision. The simplest method is to utilize the conventional ratio and regression estimator for survey estimation.

1.4 Approaches for Survey Estimation

In general, there are three broad approaches of survey sampling i.e., design-based, model-based, and model-assisted approach for the analysis of survey data. In a design-based approach, population values are assumed to be fixed. It is based on the repetition of the sampling method, i.e., selecting sample after sample from the population, calculating the value of estimate for each sample, generating a different sample error each time, and hence a distribution for these sample errors. Here, the source of variability is the sample selection method. In the model-based approach, population values are assumed to be a realization of random variables that can be characterized in terms of a statistical model. This model describes the range of possible population values that can occur and imposes a probability measure on the chance of occurrence of any particular range of values. Such models are usually based on past exposure to data from other populations very much like the one of interest as well as subject matter knowledge about how the population values ought to be distributed. In this approach, variability arises due to the distribution of values of population variables. Now many times inferences based on reliable models can be very efficient but results from mis-specified models could be disastrous under the model-based prediction approach. Design-based inferences for survey sampling, however, impose no model assumptions, and the probability sampling design is chosen by the survey sampler based on the particular survey population under study. Typically, "whatever the unknown qualities of the population," confidence intervals based on the normal theory are

asymptotically valid for large samples (Neyman, 1934). Through the use of a model-assisted technique, the strength of plausible models may be included in design-based inferences. A credible model is used to justify the building of the estimator in the model-assisted framework, but the estimator is evaluated using both the model-based and the design-based frameworks. A prediction estimator is said to be model-assisted if it meets the following criteria:

1. It is a model-unbiased prediction estimator under the underlying model.
2. Regardless of the model, it is roughly design-unbiased under the probability sampling design.

Parts I can occasionally be substituted with "roughly model-unbiased" and (ii) be substituted by "design-consistency". A stronger notion than approximate design unbiasedness is design consistency, which is the characteristic that the estimator converges in probability to the parameter of interest under the sampling design. If the finite parameter to be estimated is of order 1, the former often demands that the design-based variance goes to zero as the sample size increases.

Pioneering the work on model-assisted estimation in sample surveys Cassel *et al.* (1976) first proposed the concept of generalized regression estimator (GREG) which is the most widely used estimator under the model-assisted framework. Later Sarndal (1980) proposed the concept of regression coefficient estimation based on GREG and also showed that the proposed estimator is equally efficient as the best linear unbiased estimator. One widely used model-assisted approach to survey estimation which has gotten attention in recent years is the calibration estimation proposed by Deville *et al.* (1992). This approach originally leads to the GREG estimator of the population total under a given sampling design when the chosen distance function is the Chi-square distance (Deville *et al.*, 1992).

1.5 Calibration approach

The raking ratio estimation method of Deming *et al.* (1940), where the goal was to estimate the cell proportions in a two-way contingency table with known marginal population totals and the survey sample is taken by simple random sampling, is where the concept of calibration estimation in the presence of auxiliary information first appeared. Calibration weights were initially introduced by Huang and Fuller (1978) under the name regression weights. For complicated survey data, calibration estimates and weighting techniques were first explicitly used in household surveys. In the 1980s, a significant amount of research was done on weighting for household surveys. Later, the work of Deville *et al.* (1992) was crucial in formalizing and spreading the principles and methods of calibration weighing and estimation.

Calibration estimation is nothing but adjusting the original design weights to improve the estimates by incorporating the known population total of auxiliary variables. This is a method to improve estimation in survey sampling when auxiliary information is available. Auxiliary information is included at the estimation stage to produce efficient estimates. In this approach, survey weights are modified so that known population characteristics, in practice totals (or means), of the auxiliary variable are reproduced from the sample. Therefore, for variables in the survey correlated with the auxiliary variable, higher precision in estimates is obtained by these new weights.

There are two basic components in the construction of new calibration weights, namely a distance measure and a set of calibration constraints. The calibration weights are so chosen that they minimize a given distance measure that is the sum of chi-square type distance is minimum while satisfying constraints related to auxiliary variables. If the optimum calibrated weights do

not satisfy desired constraints of weights, then some more restrictions were added to improve the precision of the estimates.

Deville *et al.* (1993) compared alternative distance functions for constructing calibration estimators and demonstrated that various distance functions, fulfilling certain mild conditions, yield asymptotically equivalent calibration estimators. They also revealed that changes in the distance function typically have minimal impact on the variance of the calibration estimator, even when the sample size is relatively small.

Singh *et al.* (1998) developed an improved estimator of variance of the Deville *et al.* (1992) calibration estimator using higher order calibration approach. In this technique the estimator of variance of the simple calibration estimator was modified by minimizing the design weights of the variance estimator using some calibration constraints at the second order moment level. It was found that the efficiency of higher order calibration approach was better than the lower one (Deville *et al.*, 1992).

Duschene (1999) described the calibration estimators in the presence of outliers. Although his "robust calibration" method was quite effective in lowering variation, it did add some bias into the estimations.

Singh *et al.* (1999) investigated calibration approach based estimators of variance of the population total. They demonstrated how, for various sample designs, the derived calibrated estimator reduces to ratio and regression estimators.

Wu *et al.* (2001) developed a model-calibration method, suggested a unified model-assisted estimator. Under certain circumstances, the suggested model calibration estimators can reduce to the traditional calibration estimators of Deville *et al.* (1992) and can handle any linear or nonlinear working models. In this context, Chen and Sitter's (1996) pseudo empirical maximum

likelihood estimator produced an estimate that, although having positive weights, is asymptotically equal to the model-calibration estimator. The suggested estimator is based on a small number of rigid constraints and assumptions, which are typically challenging to uphold when dealing with real-world circumstances.

Tracy *et al.* (2003) presented a pair of restrictions employing first and second order moments of an auxiliary variable in order to provide calibration weights for calculating the population mean in stratified sampling. The problem of variance estimation was also considered. Double sampling was used to further the findings. Simulation research served as an illustration for the findings.

Estevao and Särndal (2003) created an effective calibrated estimator for two-stage and two-phase sampling using complicated auxiliary information (auxiliary variables at distinct stages and phases of sample design). Through the use of a linearized statistic, they were able to determine the variance and estimate of variance of the nonlinear calibration estimator.

Montanari *et al.* (2005) extended model calibration approach by taking into account more general super population models and use nonparametric methods to obtain the fitted values on which to calibrate. In order to more accurately estimate the functional connection between the survey variable and the auxiliary variables, they use neural network learning and local polynomial smoothing. Under appropriate regularity requirements, the suggested estimators are demonstrated to be design consistent.

Kott (2006) investigated the application of calibration weighting to correct for unit nonresponse and/or coverage faults. He also discovered that the generated estimator is design consistent (randomization consistent), meaning that under some conditions, the bias in the estimator's design is asymptotically minimal.

Särndal (2007) provided a thorough analysis of the calibration-related work that has been done. The estimate of a population total in direct and single-phase sampling was one of the more straightforward calibration technique applications he looked at. Then, he expanded its use to include sampling schemes and parameters of more complexity. Also covered were its uses where non-sampling error was present.

Koyuncu *et al.* (2010) proposed a calibration estimators using constraints listed in Tracy *et al.* (2003). They proposed an estimator of the population mean under stratified two-phase sampling using three calibration constraints. They add another constraint, which is sum of design weights equals to sum of calibrated weight, which is the actual bridge between the GREG estimator and traditional linear regression estimator (Singh *et al.*, 2011).

Singh *et al.* (2011) proposed a bridge between the generalized regression (GREG) estimator derived from the calibration technique of Deville *et al.* (1992) and the linear regression estimator due to Hansen *et al.* (1953). The bridge complies with Singh's (2003, 2004, 2006) observation that the sum of the calibrated weights should match the sum of the design weights. Through simulation tests for PPSWOR sampling, four distinct estimators the Ratio, GREG, Wu and Sitter (2001), and Hansen *et al.* (1953) estimators are compared. In this article, they provide the multi-auxiliary calibration estimator under the one-stage sampling design.

Rao *et al.* (2012) proposed the concept of multivariate calibration estimator for the population mean under the stratified sampling design, which incorporates information available for more than one auxiliary variable and the calibrated weights were non-negative. The problem of determining the weights with respect to the given condition of calibration on several variables was formulated and solved as a Mathematical Programming Problem (MPP).

Raman *et al.*(2013) developed calibration approach-based Hansen and Hurwitz (1946) estimator of population total for the circumstance where information on auxiliary variable was presumed known for the entire sampled units in the presence of unit nonresponse occurring in mail surveys. Expressions for the estimator of the population totals, its variance estimator were developed.

Sud *et al.* (2014) developed calibrated estimator of population total under the assumption that the auxiliary variable is negatively correlated with the study variable. The developed estimator outperformed the usual product estimator in terms of the criteria of relative bias and mean square error.

Sud *et al.* (2014) developed a regression-type estimator of the population total under the assumption that the auxiliary variable is inversely connected to the research variable. A variance estimator for the suggested estimator was developed. A higher order calibration method has also been discussed for the estimator of variance of developed estimator. A two-phase sampling strategy has been recommended when the auxiliary information was not available for all population units. The proposed estimator performed better than the existing regression estimator, according to empirical findings.

Aditya *et al.* (2016) suggested calibration-based regression type estimators of the population total with assumption that, in a two-stage sampling design, population level auxiliary information is available at primary stage unit level. The proposed estimators' variance and their estimator of variance have also developed. According to the empirical findings from the simulation tests, the suggested estimators beat the standard regression estimators under the two-stage sample design in terms of the relative bias and relative root mean square error.

Mourya *et al.* (2016) developed a calibration estimator for finite population total in two-stage sampling when the auxiliary information is available at the element level for the only selected first-stage units in the random sample. They also carried out simulation study with real data and artificial data generated through assumed regression model. The results of both simulation studies confirmed the superiority of the proposed calibration estimator over the usual estimator in two-stage sampling.

Aditya *et al.* (2017) used calibration approach proposed district level crop yield estimation. under two stage sampling design with the assumption of availability of auxiliary information at unit level only for the selected PSUs and showed that the proposed estimator performs better than the existing one through a empirical study on real survey data.

Koyuncu (2017) developed a Calibration estimator of population mean under stratified ranked set sampling design. They have used the estimator developed by Sinha *et al.* (2017) to deal with the complex auxiliary information under stratified random sampling design. Theoretical variance of the suggested estimator was discussed. Also, a simulation study was carried out to show the properties of the proposed estimator.

Nilgun Ozgul (2018) suggested a new calibration estimator for the population mean in the presence of two auxiliary variables in the stratified sampling. The theory of the novel calibration estimator is described, and the optimal calibration weights are selected utilising nonlinear constraints. The performance of the suggested calibration estimator is compared to various calibrator estimators that are already in use in a simulation study. The results demonstrate the superiority of the recommended calibration estimators over other calibrators already in use for stratified sampling.

Aditya *et al.* (2019) developed an enhanced variance estimator of the regression type estimator given by Aditya *et al.*, (2016) using a higher order calibration technique(Singh *et al.*, 1998). Additionally, a simulation study was conducted to prove the suggested estimators' empirical performance, and the findings indicate that the proposed estimator outperforms the standard estimate of variances for the regression type estimator (Aditya *et al.*, 2016).

Nilgun Ozgul (2020) considered the issue of estimating the population mean of the study variable in stratified two-phase sampling when auxiliary information is not available and proposed a new multivariate calibration technique as an alternative to the current calibration estimators. The theory of new calibration estimation is discussed under a two-phase sampling method, and the ideal weights are chosen. To evaluate the efficacy of the proposed calibration estimator with existing calibrator estimators presently in use, a simulation study is done. The results demonstrate that compared to previous calibrated population mean estimators presently in use, the proposed calibration estimator for stratified two-phase sampling is more efficient.

Alam *et al.* (2020) developed calibration estimator by taking into account the non-linear restrictions of an auxiliary variable, they established a theory of calibrated estimators of mean in simple random sampling, probability proportional to size sampling, and stratified random sampling.

Biswas *et al.* (2020) worked on Calibration Estimator in two stage sampling using double Sampling approach when study variable is inversely related to auxiliary variable. They have demonstrated through simulation study that the proposed estimator outperformed the traditional product estimator. Basak *et al.* (2021) proposed a two step calibration estimator under two stage two phase sampling design.

Clark *et al.* (2022) proposed an adaptive calibration technique for prediction of finite population totals under multivariate calibration framework where the auxiliary variables to be used in weighting were selected using sample data.

Biswas *et al.* (2023) developed a calibration estimator under two phase two stage sampling design when population level auxiliary information was not available and auxiliary variable was inversely related to the study variable. They have showed through limited simulation study that the proposed estimator was performing better than the existing estimators through the criteria of %RB and %RRMSE.

Alam *et al.* (2023) proposed a multivariate calibration estimator of the population mean by employing the multiple auxiliary variables. They introduced new variance function of the study variable in replacement to usual distance functions under the assumption of known population variance in case of Neyman allocation. In compared to the standard combined mean, combined ratio and regression estimators, the suggested estimator is proven to be more effective.

1.6 Domain Calibration estimation

Estevao and Sarndal (1999) first envisaged some important issues in the use of auxiliary information to produce design-based estimates for domains. They identified three types of design-based estimators and discussed two of these in detail. Both are defined as linear weighted sums of the observed values of the variable of interest. The first is the linear prediction estimator, which is built on a principle of model fitting and good predictions of the unobserved values of the study variable. The second is the uni-weight estimator, which applies the same weight to the study variable in the calculation of all estimates for those domains containing the respective unit. The latter approach was found to have practical advantages for large-scale productions of statistics because it does not require the calculation of different weight systems for the many

variables of interest. The second estimator was developed using the concept of calibration proposed by Deville and Sarndal (1992).

Hidiroglou and Patak (2001) in their paper entitled “Domain Estimation Using Linear Regression” introduced another concept of domain calibration estimation and its conditional properties of recognizable subsets (Rao, 1985) for various uni-stage sampling designs. The main purpose of the paper is to study the properties of a number of domain estimators of totals in the presence of auxiliary data. These properties will be established via conditioning on fixed sample sizes within each domain.

Lehtonen *et al.* (2003) examined the effect of model choice on different types of estimators for totals of domains, including small domains (small areas). In this paper they have discussed three types of estimator i.e. Synthetic, GREG, and, to a limited extent, Composite. They showed that model improvement (the transition from a weaker to a stronger model) has very different effects on the different estimator types. They also showed that the difference in accuracy between the different estimator types depends on the choice of model. For a well-specified model the difference in accuracy between Synthetic and GREG is negligible, but it can be substantial if the model is mis-specified. Synthetic then tends to be highly inaccurate.

Lehtonen *et al.* (2005) described an estimator of a total for a population subgroup or domain is with an underlying model in mind. Important features of the model include the mathematical statement of the model and the set of parameters allowed in it. They have also showed that how the features of model affect the bias and accuracy of common estimator types. They studied study two estimator types, the model dependent type and the model-assisted type. Synthetic (SYN) estimators and generalized regression (GREG) or calibration estimators are used to represent these types. Simulation results indicate that the choice of model affects the two

estimator types in very different ways. The choice between a fixed-effects model and a corresponding mixed model has a large impact on SYN, whereas the GREG estimator remains virtually unaffected.

Clement *et al.* (2014) developed an analytical approach for generating domain calibration estimator to enhance survey estimates. A mathematical programming problem (MPP) that employs the Lagrange multiplier approach to minimise the Chi-square type loss function under a number of calibration restrictions was used to express the issue of obtaining the ideal calibration weights. The ideal calibration weights adhere to the calibration constraints. The suggested domain calibration estimator outperformed the HT estimator, according to empirical research.

Hidiroglu *et al.* (2016) developed domain calibration estimators using direct and modified direct design weights under SRSWOR. Direct methods use only data within the domain where as in modified direct data from both within and outside is used for construction of the estimators.

Enang *et al.* (2019) developed an efficient class of calibration ratio estimators of domain mean in survey sampling. They proposed a new approach to domain estimation and proposed a new class of ratio estimators that is more efficient than the regression estimator and not depending on any optimality condition using the principle of calibration weightings. Some well known regression and ratio-type estimators are obtained and shown to be special members of the new class of estimators. Results of analytical study showed that the new class of estimators is superior in both efficiency and bias to all related existing estimators under review. The relative performances of the new class of estimators with a corresponding global estimator were evaluated through a simulation study.

1.2 Motivation and Objective of the Research Project

It was observed that most of the work related to domain calibration estimation for the finite population parameters was mostly restricted to only uni-stage sampling designs. But the main aim of any developed methodology was to implement the same in improvement of the estimates obtained from real life surveys. Real life surveys are generally multistage in nature and methodologies based on uni-stage designs cannot be applied directly to these survey data. Further, ignoring the survey weights will lead to inconsistent estimates of the population or domain parameters (Wu *et al.*, 2020). Hence there is an urgent need for development of the domain calibration estimation under multi-stage sampling design. Further, usually the most commonly used multistage design is two stage sampling design which was mostly used for various surveys conducted by the state and the central agencies of Government of India. Hence, the study “A Study on Domain Calibration Estimators under Two Stage Sampling Design” was proposed under the project. In this study our aim was to develop theory for estimation of domain parameters using calibration estimation technique in two stage sampling design. Hence, the project is proposed with the following objectives:

1.2.1 Immediate objectives:

- 1 To develop Calibration Estimator of Domain Total when auxiliary information is available at PSU level for each Domain.
- 2 To propose Calibration Estimator of Domain Total when auxiliary information is available at SSU level for each Domain.
- 3 To develop variance and estimator of variance of the proposed estimators under Objective 1 and 2.
- 4 To empirically evaluate the performance of the proposed estimators under objective 1 and 2.

4.2.2 Long term objectives

To use the developed estimators for producing reliable estimates of domain parameters from the real survey data.

4.3 Structure of the Project Report

This report has total four chapters. The first chapter is introductory in nature which provides an overview of Calibration estimation technique in general and motivation and objective of the research presented in this report in particular. In the next chapter we will describe the Basic concept and theory of calibration estimation by Deville and Sarndal (1992). We then define the domain calibration estimators developed in this study under two stage sampling design based on three different situations of availability of auxiliary information at the psu level and ssu level. We then define the variance and the estimate of variance of all the developed calibration estimators. In chapter 3, we will illustrate the results obtained through simulation study for all the developed estimators under two stage sampling design. Finally, Chapter 4 is devoted to concluding remarks and further research topics.

CHAPTER 2

CALIBRATION ESTIMATION TECHNIQUE

2.1 Introduction

Survey statisticians are always concerned with improvement of methods for estimation of the finite population total, mean, proportion and other parameters. The estimators which use auxiliary variables are often more accurate than the standard ones. Calibration is commonly used in survey sampling to include auxiliary information to increase the precision of the estimators of population parameter. A calibration estimator uses calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights while also respecting a set of constraints, the calibration equations. For every distance measure there is a corresponding set of calibrated weights and a calibration estimator (Deville and Särndal (1992)).

Definition: The calibration approach for estimation of finite population parameters consists of

- (a) A computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s).
- (b) The use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units.
- (c) An objective to obtain nearly design unbiased estimates as long as nonresponse and other non-sampling errors are absent.

Broad Uses of Calibration Estimation Technique

Calibration as a linear weighting method

Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign

an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies. Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horwitz and Thompson (1952). Weighting became widely accepted. Later, post stratification weighting achieved the same status. Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent; the weights depend on the observed sample.

Calibration as a systematic way to use auxiliary information

Calibration provides a systematic way to take auxiliary information into account. As Rueda *et al.* (2007) point out, “In many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation”.

Calibration to achieve consistency

Calibration is often described as “a way to get consistent estimates”. (Here “consistent” refers not to “randomization consistent” but to “consistent with known aggregates”.) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. Consistency through calibration has a broader implication than just agreement with known population auxiliary totals. Consistency can, for example, be sought with appropriately estimated totals, arising in the current survey or in other surveys.

There are three major advantages of calibration approach in survey sampling.

- I. The calibration approach leads to consistent estimates.

- II. It provides an important class of technique for the efficient combination of data sources.
- III. Calibration approach has computational advantage to calculate estimates.

The calibration approach focuses on the weights given to the units for the purpose of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. The population total is estimated by a linear estimator whose weights are as close as possible to some benchmark weights and which at the same time satisfy some calibration constraints with respect to some suitable auxiliary variables.

Consider a finite population $U = \{1, \dots, k, \dots, N\}$ consisting of N units. A sample s of size n is drawn without replacement according to a probabilistic sampling plan with inclusion probabilities $\pi_i = p_r(i \in s)$ and $\pi_{ij} = p_r(i \text{ and } j \in s)$ are assumed to be strictly positive and known. The study variable y is observed for each unit in the sample hence is known for all $i \in s$, and the values x_1, x_2, \dots, x_N are known. Let y_i be the value of the variable of interest, y , for the i^{th} population element, with which is also associated an auxiliary variable x_i . For the elements $i \in s$, observe (y_i, x_i) . The population total of auxiliary variable x , $X = \sum_{i=1}^N x_i$ is assumed to be accurately known. The objective is to estimate the population total $Y = \sum_{i=1}^N y_i$.

Deville and Sarndal (1992) used calibration on known population total X to modify the basic sampling design weights. Let the Horvitz-Thompson estimator of the population total be

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^n d_i y_i, \text{ where } d_i = \frac{1}{\pi_i}, \text{ the sampling design weight, defined as the inverse}$$

of the inclusion probability for unit i . An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design. Its variance under the sampling design is given as

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \frac{y_i \cdot y_j}{\pi_i \pi_j}$$

Now let us suppose that $\{x_i, i = 1, \dots, N\}$ is available and $X = \sum_{i=1}^N x_i$, the population total for x is known. Ideally we would like, $\sum_{i=1}^n d_i x_i = X$. But sometimes this is not true. The idea

behind calibration estimators is to find weights w_i , ($i = 1, \dots, N$) close to d_i , based on a distance function, such that, $\sum_{i=1}^n w_i x_i = X$. We wish to find weights w_i similar to d_i so as to

preserve the unbiased property of the HT estimator. Once w_i is found the calibration estimator for $Y = \sum_{i=1}^N y_i$ would be $\hat{Y}_c = \sum_{i=1}^n w_i y_i$.

Given a sample s , we want to find w_i , ($i = 1, \dots, N$) close to d_i based on a distance function $D(w, d)$ subject to the constraint equation $\sum_{i=1}^n w_i x_i = X$. The optimization problem where we

want to minimize

$$Q(w_1, \dots, w_n, \lambda) = \sum_{i=1}^n D(w_i, d_i) - \lambda \left(\sum_{i=1}^n w_i x_i - X \right) \quad \dots(2.1)$$

using the method of Lagrangian multipliers. There are various distance measures are available, some of them were,

Distance measures	$D(w, d)$
1. Chi-squared distance	$\frac{(w-d)^2}{2dq}$

2. Modified minimum entropy distance	$q^{-1}(w \log(\frac{w}{d}) - w - d)$
3. Hellinger distance	$2(\sqrt{w} - \sqrt{d})^2 / q$
4. minimum entropy distance	$q^{-1}(-d \log(\frac{w}{d}) + w - d)$
5. Modified Chi-squared distance	$\frac{(w-d)^2}{2wq}$

Here q is the tuning parameter that can be manipulated to achieve the optimum minimal of the Eq. (2.1). A simple case considered by Deville and Sarndal (1992) is the minimization of chi-square type distance function given by $\sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i q_i}$. Where, q_i is the tuning parameter. In most

of the situations, the value of $q_i = 1$. By minimizing the $\sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i q_i}$ subject to constraint

equation $\sum_{i=1}^n w_i x_i = X$ the weights w_i was obtained $w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^n d_i q_i x_i^2} \left(X - \sum_{i=1}^n d_i x_i \right)$.

Substitution of the value of w_i in $\hat{Y}_c = \sum_{i=1}^n w_i y_i$ gives

$$\begin{aligned} \hat{Y}_c &= \sum_{i=1}^n d_i y_i + \frac{\sum_{i=1}^n d_i q_i x_i y_i}{\sum_{i=1}^n d_i q_i x_i^2} \left(X - \sum_{i=1}^n d_i x_i \right) \\ &= \hat{Y}_{HT} + \hat{B} (X - \hat{X}_{HT}), \end{aligned}$$

where, $\hat{B} = \frac{\sum_{i=1}^n d_i q_i x_i y_i}{\sum_{i=1}^n d_i q_i x_i^2}$. Written in this form, we see that \hat{Y}_c is the same as the linear GREG

estimator (Cassel *et al.*, 1976). In fact, the GREG estimator is a special case of the calibration

estimator when the chosen distance function is the Chi-square distance (Deville and Sarndal, 1992). The main difference between the GREG approach and the calibration approach is in GREG approach the predicted values are generated using an assisting model whereas in calibration approach it does not depend on any assumption about the assisting model. Assisting model, an imagined relationship between study variable and auxiliary variable, can have many forms: linear, nonlinear, generalized linear, mixed (model with some fixed, some random effects), and so on. In terms of efficiency, Deville and Sarndal showed that for medium to large samples, the choice of $D(w, d)$ does not make a large impact on the variance of \hat{Y}_c . The variance of the calibration estimator was given as,

$$V(\hat{Y}_c) = V\left(\hat{Y}_{HT} + \hat{B}(X - \hat{X}_{HT})\right) \\ = \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} (d_i(y_i - Bx_i))(d_j(y_j - Bx_j))$$

As $E(\hat{B}) = B$ then B be the true population parameter and its variance will become zero. The

estimator of variance of the estimator was given as, $\hat{V}(\hat{Y}_c) = \sum_{i=1}^n \sum_{j=1}^n \frac{\Delta_{ij}}{\pi_{ij}} (w_i e_i)(w_j e_j)$. Where

$e_i = y_i - \hat{B}x_i$ and $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j)$. This technique of calibration is called as the lower level

calibration approach. Deville and Sarndal (1992) have also shown that the use of w_i in the

variance estimator makes it both design consistent and nearly model unbiased.

Domain Calibration Estimation

Hidiroglou and Patak (2001) formally introduced the concept of domain calibration estimation for equal probability without replacement sampling design. the main objectives of a sample survey is to compute estimates of means and totals of a number of characteristics associated with the units of a finite population U. The data are often used for analytic studies or analyses

of a survey. This usually involves the comparison of means and totals for subgroups of the population. Such subgroups are referred to as domains of study (Hartley, 1959). Hartley's (1959) paper is one of the first attempts to unify the theory for domains. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimators that did not make use of auxiliary information. He did, however, consider the case of the ratio estimator where population totals were known for the domains. The existence of multivariate auxiliary data raises a number of questions in the context of domain estimation. Some of those questions are as follows. What is the effect of having auxiliary information that is not known on a population basis for the given domain of interest? How do we compute valid variance estimates in the context of domain estimators that use auxiliary data? If more than one estimator is possible for point estimation and/or variance estimation, what criteria should be used to decide on how to choose the best estimator? Rao (1985) introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined after the sample has been drawn. In the context of domain estimation the number of units belonging to a particular domain is a random variable. Recognizable subsets in that context are those where the sample size is fixed within each domain. Comparison of the conditional statistical properties (i.e., bias, mean squared error) of the different estimators can then be based on these subsets. The conditioning process is that population totals are known for each domain. In the case of simple random sampling, the number of units in the population domain is assumed known.

Let the finite population $U = \{1, \dots, k, \dots, N\}$ be divided into D non-overlapping domains $U_1, \dots, U_d, \dots, U_D$ and the corresponding population size be $N_1, \dots, N_d, \dots, N_D$. Let $Y = \sum_{i=1}^N y_i$ be the population total of a characteristic of interest "y". Assume that the sampling plan, $P(s)$, is an arbitrary one with first and second order inclusion probabilities

$\pi_k = p_r(k \in s)$ and $\pi_{kl} = p_r(k \text{ and } l \in s)$. The resulting sample is denoted "s", and units in domain d that are part of s are denoted $s_d = U_d \cap s$. An estimator of the domain total

$Y_d = \sum_{k=1}^{N_d} y_k$ that does not use auxiliary data is given by,

$$\hat{Y}_{d,HT} = \sum_{s_d} a_k y_k = \sum_s a_k y_{dk},$$

where $a_k = \frac{1}{\pi_k}$, the survey design weight,

$$y_{dk} = \begin{cases} y_k, & \text{if } k \in s_d \\ 0, & \text{otherwise} \end{cases}.$$

They assumed that, auxiliary information in the form of a p -dimensional vector \mathbf{x} may be available at different levels of aggregation. They have assumed that, $X_d = \sum_{k=1}^{N_d} x_k$ is completely known and will be estimated as $\hat{X}_{d,HT} = \sum_{k=1}^{n_d} a_k x_k$. Now the new w_k weight is determined using calibration approach by minimizing the chi-square type distance function with respect to the calibration constraint $X_d = \sum_{k=1}^{n_d} w_k x_k$ and the putting the value of the calibrated estimator of the domain total will be given as,

$$\hat{Y}_{d,cal} = \hat{Y}_{d,HT} + (X_d - \hat{X}_{d,HT}) \hat{B}_d \quad (2.2)$$

where, $\hat{B}_d = \left(\frac{a_k x_k^2}{q_k} \right)^{-1} \sum_{s_d} \frac{a_k x_k y_k}{q_k}$

Additionally, with this estimator they have also proposed a Hajek type domain calibration estimators. According to Sarndal, Swenson, and Wretman (1992, p. 182), the Hajek type estimator is ‘usually the better estimator’ comparing to the Horvitz-Thompson estimator and this is the main reason they have proposed the Hajek type extension of the above estimator.

Clement *et al.* (2014) provided the design consistent and model unbiased estimate of the variance of the domain calibration estimator proposed by Hidiroglou and Patak (2001). The estimator of variance of the estimator in equation (2.2) was given as,

$$\hat{V}(\hat{Y}_{d,cal}) = \sum_{k=1}^{n_d} \sum_{l=1}^{n_d} \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_{dk})(w_l e_{dl})$$

(2.3)

where, $e_{dk} = y_d - x_d \hat{B}_d$

2.2 Proposed Calibration estimators under two stage sampling design

In many medium to large scale surveys, it is very often the case that we do not have a sampling frame. In some cases, the population could be spread over a wide area entailing very high travel expenses for the personal interviewers and efficient supervision of the field work can be difficult. In these situations, we prefer to use multistage sampling designs. Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario one can ask to estimate the mixture wise crop statistics which is a common case of domain estimation. In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of finite population total or mean or distribution function. In the simplest settings, ratio and regression estimators incorporate known finite population parameters of auxiliary variables. The Calibration

Approach (Deville and Sarndal, 1992) is one of the techniques widely used for making efficient use of auxiliary information in survey estimation by providing new set of weights by modifying the sampling design weights using auxiliary information. Now to address the problem of domain estimation and to improve the domain specific estimators under two stage sampling design scenario a domain calibration estimator is developed under two stage sampling design. Under two stage sampling design frame work the proposed estimators were developed with the assumption that there was availability of auxiliary informations both at psu and ssu level. We have considered the simple case where information on only one auxiliary variable is available.

Let, the population of elements $U = \{1, \dots, k, \dots, N\}$ is partitioned into clusters, $U_1, U_2, \dots, U_i, \dots, U_{N_i}$. They are also called the primary stage units (psus) when there are two stages of selection.

The size of U_i is denoted as N_i . We have $U = \bigcup_{i=1}^{N_i} U_i$ and $N = \sum_{i=1}^{N_i} N_i$. We are considering the direct domain Estimation Scenario in this study and all are planned domains (Hidiroglou, 2016) i.e. sufficient sample size exists for direct estimates. Further, we assume that there are D domains $U_1, \dots, U_d, \dots, U_D$ ($d=1, 2, \dots, D$). Let N_{id} psus among N_i psus contain units belonging to d^{th} domain. Further, let N_{id} units out of N_i units in i^{th} psu fall in the d^{th} domain. Let y_{ikd} be the value of the study character under consideration corresponding to k^{th} ssu in the i^{th} psu pertaining to d^{th} domain.

The total population size is,

$$N = \sum_{d=1}^D \sum_{i=1}^{N_{id}} N_{id}$$

The domain total for d-th domain will be given as,

$$Y_d = \sum_{i=1}^{N_{id}} \sum_{k=1}^{N_{id}} y_{ikd}$$

The population total for all the domains will be defined as,

$$Y = \sum_{d=1}^D Y_d$$

Now a probability sample of n_I is drawn from N_I at psu level and a sample of n_i is drawn from N_i at ssu level where n_{Id} psus at out of n_I psus and n_{id} ssus out n_i SSUS falls in the d -th domain.

Following, Sarndal *et al.* (1992) and Enang *et al.* (2014) the first and second order inclusion probabilities will be defined as,

$$\pi_{i_i} = \Pr(i \in s_I),$$

$$\pi_{Iij} = \begin{cases} \Pr(i \& j \in s_I), i \text{ and } j \text{ belongs to different psus} \\ \pi_{i_i}, i \text{ and } j \text{ belongs to same psus} \end{cases}$$

and

$$\pi_{k/i} = \Pr(k \in s_i | i \in s_I) \text{ and}$$

$$\pi_{kl/i} = \begin{cases} \Pr(k \& l \in s_i | i \in s_I), k \text{ and } l \text{ are different} \\ \pi_{k/i}, k \text{ and } l \text{ are same} \end{cases}$$

Now at psu level, now the HT estimator of the domain total at d -th domain will be given as,

$$\hat{Y}_{d2st} = \sum_{i=1}^{n_{Id}} \frac{1}{\pi_{i_i}} \sum_{k=1}^{n_{id}} \frac{y_{ikd}}{\pi_{k/i}} = \sum_{i=1}^{n_{Id}} a_{i_i} \hat{t}_{yi\pi d}$$

Where, $a_{i_i} = \frac{1}{\pi_{i_i}}$ is the design weight at psu level and $\hat{t}_{yi\pi d} = \sum_{k=1}^{n_{id}} \frac{y_{ikd}}{\pi_{k/i}}$ =HT estimator of the psu total at d -th domain.

Three cases of availability of auxiliary information was considered for the construction of the estimator, which were,

Case 1. The domain level auxiliary information (z_{id}) is available at the psu level i.e. for national surveys for certain establishments say hospitals (study variable) under each village, information on each village can be easily obtained and used as a auxiliary variable at psu level.

Case 2. Domain level auxiliary information (x_{kd}) is available only for the selected psus.

There is another situation of availability of auxiliary information at SSU level (Wu *et al.*, 2021) which is availability of domain level complete auxiliary information for all the units. This condition of availability of auxiliary information is very rare and usually very hard to find even for the entire population and hence it is difficult to obtain for the situation of domain estimation. Hence this situation is not considered in this study.

Case1. Let us assume that, domain level auxiliary information z_{id} was available at the psu level and the correlation between the study variable and the auxiliary variable was positive and the value of z_{id} was observed for all the sampled clusters under d -th domain and a correct value of

$Z_d = \sum_{i=1}^{N_{ld}} z_{id}$ was accurately known. Now following the concept of calibration ideally we have,

$Z_d = \sum_{i=1}^{n_{ld}} a_{li} z_{id}$. Now in calibration our aim is to find out a weight w_{li} such that, $Z_d = \sum_{i=1}^{n_{ld}} w_{li} z_{id}$

. Now if we put w_{li} in proposed estimator we get the calibration estimator of domain total which will be,

$$\hat{Y}_{d2stc}^c = \sum_{i=1}^{n_{ld}} w_{li} \hat{y}_{i\pi d}$$

Now by minimizing the chi-square type distance function $\sum_{i=1}^{n_{ld}} \frac{(w_{li} - a_{li})^2}{a_{li} q_{li}}$ subject to the

constraint, $\sum_{i=1}^{n_{ld}} w_{li} z_{id} = Z_d$, the new calibrated weight will be given as,

$$w_{li} = a_{li} + a_{li} q_{li} z_{id} \left[\frac{Z_d - \sum_{i=1}^{n_{ld}} a_{li} z_{id}}{\sum_{i=1}^{n_{ld}} a_{li} q_{li} z_{id}^2} \right]$$

Assuming $q_{li} = 1$, the calibration estimator will be given as,

$$\hat{Y}_{d2stc}^c = \sum_{i=1}^{n_{Id}} w_{li} \hat{t}_{yi\pi d} = \sum_{i=1}^{n_{Id}} a_{li} \hat{t}_{yi\pi d} + \sum_{i=1}^{n_{Id}} a_{li} z_{id} \hat{t}_{yi\pi d} \left[\frac{Z_d - \sum_{i=1}^{n_{Id}} a_{li} z_{id}}{\sum_{i=1}^{n_{Id}} a_{li} z_{id}^2} \right]$$

Hence, $\hat{Y}_{d2stc}^c = \hat{Y}_{dHT} + \hat{\beta} [Z_d - \hat{Z}_{dHT}]$

(2.3)

Where, $\hat{\beta} = \frac{\sum_{i=1}^{n_{Id}} a_{li} z_{id} \hat{t}_{yi\pi d}}{\sum_{i=1}^{n_{Id}} a_{li} z_{id}^2}$.

Now, from equation (2.3) we have a GREG estimator of the domain total under two stage sampling design when PSU level auxiliary information is available similar to the one proposed by Hidiroglou and Patak (2001) for uni-stage sampling design in d -th domain. This estimator is unconditionally unbiased (Hidiroglou and Patak, 2001).

It can be seen that, the estimator in Eq. (2.3) takes the form of regression estimator for the domain total under two stage sampling design. Now if we put $q_{li} = \frac{1}{z_{id}}$, the estimator in Eq.(2.3)

take the form,

$$\hat{Y}_{d2stc}^c = \frac{\sum_{i=1}^{n_{Id}} a_{li} \hat{t}_{yi\pi d} \sum_{i=1}^{N_{Id}} z_{id}}{\sum_{i=1}^{n_{Id}} a_{li} z_{id}} .$$

(2.3.1)

The estimator given in Eq. (2.3.1) was a domain calibration ratio estimator under two stage sampling design. The theoretical bias of this ratio estimator obtained through Taylor series linearization technique was given as,

$$Bias(\hat{Y}_{d2stc}^c) = \frac{1}{\sum_{i=1}^{N_{Id}} z_{id}} \left[R \sum_{i=1}^{N_{Id}} \sum_{j=1}^{N_{Id}} \Delta_{lij} \frac{z_{id}}{\pi_{li}} \frac{z_{jd}}{\pi_{lj}} - \sum_{i=1}^{N_{Id}} \sum_{j=1}^{N_{Id}} \Delta_{lij} \frac{t_{yd}^c}{\pi_{li}} \frac{z_{jd}}{\pi_{lj}} \right]$$

$$\text{where, } R = \frac{\sum_{i=1}^{N_{id}} y_{(c)id}}{\sum_{i=1}^{N_{id}} z_{id}} \text{ and } t_{yd}^c = \sum_{k=1}^{N_{id}} y_{ikd}.$$

Following Sarndal *et al.* (1992) this estimator in Eq. (2.3) can also be written as,

$$\hat{Y}_{d2stc} = \sum_{i=1}^{n_{id}} w_{i'} \hat{t}_{yi\pi d} = \sum_{i=1}^{n_{id}} a_{i'} g_{is'} \hat{t}_{yi\pi d}$$

Now to find out the approximate variance of the proposed estimator, we have used the Taylor series expansion using partial derivatives as Sarndal *et al.* (1992) is,

$$V(\hat{Y}_{d2stc}) = \sum_{i=1}^{N_{id}} \sum_{\substack{j=1 \\ i \neq j}}^{N_{id}} \Delta_{ij} \frac{e_{i'}}{\pi_{i'}} \frac{e_{j'}}{\pi_{j'}} + \sum_{i=1}^{N_{id}} \frac{1}{\pi_{i'}} \sum_{k=1}^{N_{id}} \sum_{\substack{l=1 \\ k \neq l}}^{N_{id}} \Delta_{kl/i} \frac{y_{idk}}{\pi_{k/i}} \frac{y_{idl}}{\pi_{l/i}},$$

Where,

$$e_{i'} = t_{yid} - \beta z_{id}, \quad \Delta_{ij} = (\pi_{ij} - \pi_{i'}\pi_{j'}), \quad \Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i}, \quad t_{yid} = \sum_{k=1}^{N_{id}} y_{idk}, \quad \text{and } \beta = \frac{\sum_{i=1}^{N_{id}} a_{i'} z_{id} t_{yid}}{\sum_{i=1}^{N_{id}} a_{i'} z_{id}^2}.$$

Following the model assisted survey sampling approach by Sarndal *et al.* (1992) and Wu *et al.* (2020) the Yates-Grundy form of estimator of variance of the calibration estimator given in Eq. (2.8) was given by,

$$\hat{V}(\hat{Y}_{d2stc}) = \frac{1}{2} \sum_{i=1}^{n_{id}} \sum_{j=1}^{n_{id}} d_{ij} (w_{i'} u_{id} - w_{j'} u_{jd})^2 + \frac{1}{2} \sum_{i=1}^{n_{id}} \frac{g_{is'}^2}{\pi_{i'}^2} \sum_{k=1}^{n_{id}} \sum_{l=1}^{n_{id}} d_{kl/i} \left(\frac{y_{idk}}{\pi_{k/i}} - \frac{y_{idl}}{\pi_{l/i}} \right)^2, \quad (2.4)$$

$$\text{where, } u_{id} = t_{yd}^c - \hat{\beta} z_{id}, \quad d_{ij} = \frac{(\pi_{i'}\pi_{j'} - \pi_{ij})}{\pi_{ij}}, \quad d_{kl/i} = \frac{(\pi_{k/i}\pi_{l/i} - \pi_{kl/i})}{\pi_{kl/i}} \quad \text{and } \hat{\beta} = \frac{\sum_{i=1}^{n_{id}} a_{i'} z_{id} \hat{t}_{yi\pi d}}{\sum_{i=1}^{n_{id}} a_{i'} z_{id}^2}.$$

Case 2. Domain level auxiliary information was available at the unit (ssu) level only for the selected psus i.e. the auxiliary information x_{kd} was known for all elements $k \in s$ while correct

value of $\sum_{k=1}^{N_{id}} x_{kd}$ was available for each sampled psu's and the correlation between the study

variable and the auxiliary variable was positive.

The simple HT estimator of the population total in this case will be,

$$\hat{t}_{HT} = \sum_{i=1}^{n_{id}} a_{i1} \sum_{k=1}^{n_{id}} a_{k/i} y_{kd} = \sum_{k=1}^{n_{id}} a_k y_{kd}$$

The proposed calibration estimator of the population total in this case is given as,

$$\hat{t}_{y\pi d}^c = \sum_{i=1}^{n_{id}} a_{i1} \sum_{k=1}^{n_{id}} w_k^* y_{kd} \quad (2.5)$$

w_k^* was the calibrated weight corresponding to the design weight $a_{k/i}$ in this situation. Here, we minimize the chi-square type distance function using Lagrangian multiplier technique as described in the earlier cases and obtain the calibrated weight. Here, we minimize,

$$\sum_{k=1}^{n_{id}} \frac{(w_k^* - a_{k/i})^2}{a_{k/i} q_k^*} \text{ such that } \sum_{k=1}^{n_{id}} w_k^* x_{kd} = \sum_{k=1}^{N_{id}} x_{kd};$$

Hence, the calibrated weight will be given as,

$$w_k^* = a_{k/i} + \frac{a_{k/i} q_k^* x_{kd}}{\sum_{k=1}^{n_{id}} a_{k/i} q_k^* x_{kd}^2} \left(\sum_{k=1}^{N_{id}} x_{kd} - \sum_{k=1}^{n_{id}} a_{k/i} x_{kd} \right).$$

After considering $q_k^* = 1$ the estimator becomes,

$$\begin{aligned} \hat{t}_{y\pi d}^c &= \sum_{i=1}^{n_{id}} a_{i1} \left[\sum_{k=1}^{n_{id}} a_{k/i} y_{kd} + \sum_{k=1}^{n_{id}} \frac{a_{k/i} x_{kd} y_{kd}}{\sum_{k=1}^{n_{id}} a_{k/i} x_{kd}^2} \left(\sum_{k=1}^{N_{id}} x_{kd} - \sum_{k=1}^{n_{id}} a_{k/i} x_{kd} \right) \right] \quad (2.6) \\ &= \sum_{i=1}^{n_{id}} a_{i1} \left[\sum_{k=1}^{n_{id}} a_{k/i} g_{ks}^d y_{kd} \right] \end{aligned}$$

Now, considering, $q_k^* = \frac{1}{x_{kd}}$ gives,

$$\hat{t}_{y\pi d}^r = \sum_{i=1}^{n_{id}} a_{li} \left[\frac{\sum_{k=1}^{n_{id}} a_{k/i} y_{kd} \left(\sum_{k=1}^{N_{id}} x_{kd} \right)}{\sum_{k=1}^{n_{id}} a_{k/i} x_{kd}} \right]$$

The above estimator takes the form of a ratio estimator under this condition.

The Approximate variance of the proposed estimator under **Case 2** was obtained by first order Taylor series linearization technique and was given by

$$V(\hat{t}_{y\pi d}^{**}) = \sum_{i=1}^{N_{id}} \sum_{j=1}^{N_{id}} \Delta_{ij} \frac{t_{y_{id}}}{\pi_{li}} \frac{t_{y_{jd}}}{\pi_{lj}} + \sum_{i=1}^{N_{id}} \frac{1}{\pi_{li}} \sum_{k=1}^{N_{id}} \sum_{l=1}^{N_{id}} \Delta_{kl/i} \frac{E_k''}{\pi_{k/i}} \frac{E_l''}{\pi_{l/i}} \quad (2.7)$$

where, $E_k'' = y_{kd} - \beta'' x_{kd}$, $t_{y_{id}} = \sum_{k=1}^{N_{id}} y_{kd}$, $\Delta_{ij} = (\pi_{lij} - \pi_{li}\pi_{lj})$, $\Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i}$ and

$$\beta'' = \frac{\sum_{k=1}^N y_{kd} x_{kd}}{\sum_{i=1}^N x_{kd}^2}.$$

To get the approximate variance expression we have approximated the $g_{ks}^d = 1$.

The approximate form of estimator of variance of the calibration estimator was given by,

$$\hat{V}_{YG}(\hat{t}_{y\pi d}^{**}) = \frac{1}{2} \sum_{i=1}^{n_{id}} \sum_{j=1}^{n_{id}} d_{ij} \left(\frac{\hat{t}_{y_{i\pi d}}}{\pi_{li}} - \frac{\hat{t}_{y_{j\pi d}}}{\pi_{lj}} \right)^2 + \frac{1}{2} \sum_{j=1}^{n_{id}} \frac{1}{\pi_{li}^2} \sum_{k=1}^{n_{id}} \sum_{l=1}^{n_{id}} d_{kl/i} (w_k^* e_{ks} - w_l^* e_{ls})^2, \quad (2.8)$$

where, $e_{ks} = y_{kd} - \hat{\beta}'' x_{kd}$, $d_{ij} = \frac{(\pi_{li}\pi_{lj} - \pi_{lij})}{\pi_{lij}}$, $d_{kl/i} = \frac{(\pi_{k/i}\pi_{l/i} - \pi_{kl/i})}{\pi_{kl/i}}$, $\hat{t}_{y_{i\pi d}} = \sum_{k=1}^{n_{id}} a_{k/i} y_{kd}$ and

$$\hat{\beta}'' = \frac{\sum_{i=1}^{n_{sd}} a_k y_{kd} x_{kd}}{\sum_{i=1}^{n_{sd}} a_k x_{kd}^2}.$$

CHAPTER 3

EMPIRICAL STUDY

3.1 Empirical Evaluation

In this chapter, we report the results from simulation studies that aim at assessing the performance of the developed domain calibration estimators under two stage sampling design.

In this study we have considered the case of two stage sampling where sample selection at each stage is governed by equal probability without replacement sampling design (SRSWOR). We have compared our proposed estimator with the domain level Horvitz-Thompson estimator under two stage sampling design as given in Sarndal *et al.* (1992) for both the situations when domain level auxiliary information was available at the PSU and the SSU level.

sizes were drawn to conduct the simulation study. In the case when domain level auxiliary A population was generated using model based simulation of size 20000 which consist of $N_I=400$ psus and $N_i=50$ SSUs. Both the study variable and auxiliary variable was developed as follows,

First the auxiliary variable z_i is generated independently from a normal distribution with mean 5 and variance $\sigma_x^2 = 3$ i.e. $z_i : N(5, \sigma_x^2)$.

$$y_i = \alpha_0 + \beta_1 z_i + e_{i1}; \quad i = 1, 2, \dots, N$$

We fixed $\alpha_0 = 70$, $\beta_1 = 4$ and $e_{i1} \sim N(0, 1)$

The population then was divided in three domains i.e.

$$N_{I1} = 120, N_{i1} = 50 \rightarrow \text{Domain 1}$$

$$N_{I2} = 80, N_{i2} = 50 \rightarrow \text{Domain 2}$$

$$N_{I3} = 200, N_{i3} = 50 \rightarrow \text{Domain 3. } (d=1, 2, 3)$$

The correlation between the study variable and the auxiliary variables were considered to be around 0.82.

Now from each domain various combinations of PSU and SSU sample information was available at the cluster level (Case 1, Eq. 2.3), we have considered the following combinations of sample sizes in each of the three domains as given in Table 1.

Table 1. Various Combinations of Sample Sizes for Case 1

Set	PSUs	SSUs
1	30	10
2	30	15
3	30	20
4	40	20
5	40	30
6	40	40
7	50	20
8	50	30
9	50	40

For each case, a simple random sample without replacement (SRSWOR) sample of size n_{id} psus were first taken from N_{id} psus and then from each psu a sample of n_{id} ssus were drawn by SRSWOR. Subsequently, the estimation of domain total was carried out. In particular, we repeated the simulation process $R= 10000$ times and calculated the estimates of domain total.

For Case 2 and Eq. 2.6, when domain level auxiliary information was available at the SSU level, a separate population was generated using model based simulation of size $N=20000$ which consist of $N_I=400$ psus and $N_i=50$ SSUs. Both the study variable and auxiliary variable was developed as follows,

First the auxiliary variable x_k is generated independently from a normal distribution with mean 8 and variance $\sigma_x^2 = 2.0$ i.e. $x_k : N(8, \sigma_x^2)$.

Then the study variable was generated using the model,

$$y_i = \alpha_{0k} + \beta_{1k}x_k + e_k; i = 1, 2, \dots, N$$

we fixed, $\alpha_{0k} = 40$, $\beta_{1k} = 10$ and $e_k \sim N(0, 1)$

The population then was divided in four domains i.e.

$$N_{I1} = 120, N_{i1} = 50 \rightarrow \text{Domain 1}$$

$$N_{I2} = 130, N_{i2} = 50 \rightarrow \text{Domain 2}$$

$$N_{I3} = 150, N_{i3} = 50 \rightarrow \text{Domain 3 (Here, } d=1, 2, 3)$$

The correlation between the study variable and the auxiliary variables were considered to be around 0.85. The various combinations of sample sizes considered in each domain were as given in Table 2.

Table 2. Various Combinations of Sample Sizes for Case 2

Set	PSUs	SSUs
1	20	10
2	20	15
3	20	20
4	30	10
5	30	15
6	30	20
7	40	20
8	40	30
9	40	40

3.2 Performance Measures

Developed estimators will be evaluated on the basis of two measures viz. percentage Relative Bias (%RB) and percentage Relative Root Mean Squared Error (%RRMSE). The formula of Relative Bias and Relative Root Mean Squared Error of any estimator of the population parameter θ are given by

$$RB(\hat{\theta}) = \frac{1}{S} \sum_{i=1}^S \left(\frac{\hat{\theta}_i - \theta}{\theta} \right) \times 100,$$

$$RRMSE(\hat{\theta}) = \frac{1}{\theta} \sqrt{\frac{1}{S} \sum_{i=1}^S (\hat{\theta}_i - \theta)^2} \times 100$$

where, $\hat{\theta}_i$ are the value of the estimator generated through simulation study and θ is the overall population total for the character under study.

The following tables contain the results obtained for each of the cases considered under simulation study for lower level calibration approach.

Table 3 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the PSU level for domain 1.

Table 4 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the PSU level for domain 2.

Table 5 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the PSU level for domain 3.

Table 6 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the SSU level for domain 1.

Table 7 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the SSU level for domain 2.

Table 8 contains the %RB and %RRMSE for the proposed estimator w.r.t. the Horvitz-Thompson estimator for estimating the domain total under two stage sampling design when domain level auxiliary information is available at the SSU level for domain 3.

3.3 Results

The results of the empirical evaluation were as follows

Table 3. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 1, for domain level auxiliary information was available at PSU level.

Set			%RB		%RRMSE	
	n_{ld}	n_{id}	\hat{Y}_{d2st}	\hat{Y}_{d2st}^c	\hat{Y}_{d2st}	\hat{Y}_{d2st}
1	30	10	-0.009	0.021	14.407	10.401
2	30	15	0.009	0.017	14.348	10.341
3	30	20	0.008	0.015	14.257	09.253
4	40	10	-0.009	0.019	14.252	10.348
5	40	15	0.009	0.014	14.252	09.237
6	40	20	0.007	0.011	14.224	09.216
7	50	10	0.009	0.013	14.311	10.301
8	50	15	0.007	0.012	14.256	09.016
9	50	20	0.006	0.011	14.229	09.009

Table 4. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 2, for domain level auxiliary information was available at PSU level.

Set	n_{ld}	n_{id}	%RB		%RRMSE	
			\hat{Y}_{d2st}	\hat{Y}_{d2st}^c	\hat{Y}_{d2st}	\hat{Y}_{d2st}
1	30	10	-0.009	0.020	14.394	11.391
2	30	15	0.009	0.015	14.338	10.331
3	30	20	0.008	0.013	14.257	10.253
4	40	10	-0.009	0.019	14.252	10.338
5	40	15	0.009	0.014	14.265	10.236
6	40	20	0.007	0.011	14.267	10.228
7	50	10	0.009	0.013	14.244	10.291
8	50	15	0.007	0.012	14.226	09.214
9	50	20	0.006	0.011	14.219	09.258

Table 5. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 3, for domain level auxiliary information was available at PSU level.

Set	n_{ld}	n_{id}	%RB		%RRMSE	
			\hat{Y}_{d2st}	\hat{Y}_{d2st}^c	\hat{Y}_{d2st}	\hat{Y}_{d2st}
1	30	10	-0.010	0.021	15.291	12.048
2	30	15	-0.009	0.015	15.238	12.309
3	30	20	0.007	0.013	14.257	11.853
4	40	10	-0.009	0.019	15.252	12.331
5	40	15	0.009	0.013	14.265	11.836
6	40	20	0.007	0.011	14.267	10.100
7	50	10	0.009	0.014	14.244	12.391
8	50	15	0.007	0.012	14.226	11.214
9	50	20	0.006	0.011	14.202	10.002

Table 6. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 1, for domain level auxiliary information was available at SSU level.

Set	n_{Id}	n_{id}	%RB		%RRMSE	
			\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$	\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$
1	20	10	-0.008	0.019	16.474	14.114
2	20	15	-0.007	0.017	15.452	13.872
3	20	20	-0.007	0.015	15.312	13.303
4	30	10	0.007	0.016	15.435	13.806
5	30	15	0.006	0.016	15.312	13.302
6	30	20	0.006	0.014	14.271	12.265
7	40	10	-0.006	0.013	15.375	13.312
8	40	15	0.006	0.013	14.223	12.211
9	40	20	0.005	0.011	13.204	10.197

Table 7. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 2, for domain level auxiliary information was available at SSU level.

Set	n_{Id}	n_{id}	%RB		%RRMSE	
			\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$	\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$
1	20	10	-0.008	0.015	15.273	12.786
2	20	15	-0.007	0.014	15.252	12.672
3	20	20	0.007	0.014	15.117	12.433
4	30	10	0.007	0.014	15.255	12.606
5	30	15	0.006	0.013	15.111	12.402
6	30	20	0.006	0.012	14.358	12.195
7	40	10	0.006	0.013	15.116	12.435
8	40	15	0.006	0.012	14.323	12.001
9	40	20	0.005	0.011	13.387	11.181

Table 8. %RB and %RRMSE of the proposed estimator against the HT estimator for domain 3, for domain level auxiliary information was available at SSU level.

Set	n_{ld}	n_{id}	%RB		%RRMSE	
			\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$	\hat{t}_{HT}	$\hat{t}_{y\pi d}^c$
1	20	10	0.007	0.013	15.115	12.998
2	20	15	0.007	0.013	14.852	12.872
3	20	20	0.006	0.012	14.598	12.617
4	30	10	0.007	0.013	14.855	12.806
5	30	15	0.006	0.012	15.511	12.602
6	30	20	0.006	0.011	14.258	12.085
7	40	10	0.006	0.013	14.516	12.635
8	40	15	0.006	0.012	14.233	11.901
9	40	20	0.005	0.010	13.814	11.208

3.4 Discussion

In this simulation study we have made comparison among the domain level Horvitz Thompson estimator (Sarndal et al., 1992) with the proposed domain regression type calibration estimators under two stage sampling design. Two cases of availability of auxiliary information was considered i.e.

Case 1. Domain level auxiliary information was available at the PSU level

Case 2. Domain level auxiliary information was available at the SSU level only for the selected PSUs.

In both the cases, a regression type and a ratio type estimator was proposed using the calibration estimation technique. We have considered the regression type estimator for the empirical evaluation as regression type estimators are almost unbiased under certain conditions and assumptions rather than the ratio estimators. The Horvitz-Thompson estimator was considered under the simulation study as no other published estimators under domains were available

under two stage sampling design to suit the conditions laid in this study. Under the empirical evaluation, three random domains were created from an artificial population generated using R software. R-code was developed for the simulation study and around 10000 iterations were run to find out the results based on %RB and %RRMSE of the estimators for comparison. The results as found in Table 3,4 and 5 depicts the condition of availability of auxiliary information at the PSU level where as the table 6, 7 and 8 depicts the condition of availability of auxiliary information at the SSU level.

From Table 3, 4 and 5 it can be seen that, for domain 1 ($N_{11} = 120$, $N_{i1} = 50$), domain 2 ($N_{12} = 80$, $N_{i2} = 50$) and domain 3 ($N_{13} = 200$, $N_{i3} = 50$) respectively, the value of %RB of the proposed estimator was almost similar with the domain level unbiased Horvitz Thompson (HT) estimator under two stage sampling design when selection of the units was done using SRSWOR. In all the cases the proposed regression estimator under Case 1 (domain level auxiliary information was available at the PSU level) is having slightly more %RB than the HT estimator. The reason behind that, ratio and regression estimators were usually biased estimators. Under certain specific conditions i.e. the study variable and the auxiliary variable is perfectly linearly related with pearson's correlation coefficient value as 1 and up to first order tailor series approximation, the regression estimator will become an unbiased estimator of the population/domain parameters. But in our case we have considered the correlation between the study and auxiliary variable as moderate ($=0.82$) and the estimator under consideration is a model assisted estimator i.e. Generalized Regression Estimator (GREG) which is different from the classical regression estimator (Hansen et al., 1953) by definition. Generally, the GREG estimators were consistent but not unbiased with respect to the population parameters. Hence, due to these reasons, our proposed estimator is performing almost at par with the usual HT estimator of the domain total from the criteria of %RB when population level auxiliary information is available at the PSU level under two stage sampling design scenario.

Further, from Table 3, for domain 1 ($N_{II}=120, N_{i1}=50$), it can be seen that, the proposed GREG estimator of the domain total, is performing much better than the usual HT estimator from the criteria of %RRMSE. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 09.009 to 10.401 for the sample sizes $n_{II}=50, n_{i1}=20$ to $n_{II}=30, n_{i1}=10$ respectively while the HT estimator gives %RRMSE in the range 14.229 to 14.407 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 1 when PSU level auxiliary information is available at the domain level.

From Table 4, for domain 2 ($N_{II}=80, N_{i1}=50$), it can be seen that, the proposed GREG estimator of the domain total, is also performing much better than the usual HT estimator from the criteria of %RRMSE. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 09.258 to 10.391 for the sample sizes $n_{II}=50, n_{i1}=20$ to $n_{II}=30, n_{i1}=10$ respectively while the HT estimator gives %RRMSE in the range 14.219 to 14.394 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 2 when PSU level auxiliary information is available at the domain level.

From Table 5, for domain 3 ($N_{II} = 200$, $N_{i1} = 50$), it can be seen that, the proposed GREG estimator of the domain total, is also performing better than the usual HT estimator from the criteria of %RRMSE. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 10.002 to 12.048 for the sample sizes $n_{II} = 50$, $n_{i1} = 20$ to $n_{II} = 30$, $n_{i1} = 10$ respectively while the HT estimator gives %RRMSE in the range 14.202 to 15.291 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 3 when PSU level auxiliary information is available at the domain level.

From Table 6, 7 and 8 it can be seen that, for domain 1 ($N_{I1} = 120$, $N_{i1} = 50$), domain 2 ($N_{I2} = 130$, $N_{i2} = 50$) and domain 3 ($N_{I3} = 150$, $N_{i3} = 50$) respectively, the value of %RB of the proposed estimator was almost similar with the domain level unbiased HT estimator under two stage sampling design when selection of the units was done using SRSWOR. In all the cases the proposed GREG estimator under Case 2 (domain specific auxiliary information is available at the SSU level only for the selected PSUs) is having little more %RB than the HT estimator. The reason behind that, ratio and regression estimators were usually biased estimators. Under certain specific conditions i.e. the study variable and the auxiliary variable is perfectly linearly related with pearson's correlation coefficient value as 1 and up to first order Taylor series approximation, the regression estimator will become an unbiased estimator of the population/domain parameters. But in our case we have considered the correlation between the

study and auxiliary variable as moderate ($\rho=0.85$) and the estimator under consideration is a model assisted estimator i.e. Generalized Regression Estimator (GREG) which is different from the classical regression estimator (Hansen *et al.*, 1953) by definition. Generally, the GREG estimators were consistent but not unbiased with respect to the population parameters. Hence, due to these reasons, our proposed estimator is performing almost at par with the usual HT estimator of the domain total from the criteria of %RB when population level auxiliary information is available at the PSU level under two stage sampling design scenario.

From Table 6, for domain 1 ($N_{II} =120, N_{iI} =50$), it can be seen that, the proposed GREG estimator of the domain total under Case 2, performing consistently better than the usual HT estimator from the criteria of %RRMSE for all sample sizes. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 10.197 to 14.114 for the sample sizes $n_{II}=40, n_{iI} =20$ to $n_{II}=20, n_{iI} =10$ respectively while the HT estimator gives %RRMSE in the range 13.204 to 16.474 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 1 when SSU level auxiliary information is available at the domain level.

From Table 7, for domain 2 ($N_{II} =130, N_{iI} =50$), it can be seen that, the proposed GREG estimator of the domain total under Case 2, performing consistently better than the usual HT estimator from the criteria of %RRMSE for all sample sizes. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 11.181 to 12.786 for the sample

sizes $n_{II}=40$, $n_{i1}=20$ to $n_{II}=20$, $n_{i1}=10$ respectively while the HT estimator gives %RRMSE in the range 13.387 to 15.273 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 2 when SSU level auxiliary information is available at the domain level.

From Table 8, for domain 3 ($N_{II}=150$, $N_{i1}=50$), it can be seen that, the proposed GREG estimator of the domain total under Case 2, performing consistently better than the usual HT estimator from the criteria of %RRMSE for all sample sizes. The %RRMSE of the proposed GREG estimator of the domain total varies from minimum of 11.208 to 12.998 for the sample sizes $n_{II}=40$, $n_{i1}=20$ to $n_{II}=20$, $n_{i1}=10$ respectively while the HT estimator gives %RRMSE in the range 13.814 to 15.115 for the same sample sizes. There is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase from 10 to 20. Further, it can also be seen that, when SSU level sample size is kept fixed there was improvement in the %RRMSE of the proposed estimator when PSU level sample size increase from 30 to 50. Hence, we can conclude that the proposed estimator is performing much better than the existing HT estimator for domain 3 when SSU level auxiliary information is available at the domain level.

3.5 Summary of the Major Findings

A close perusal of Tables 3, 4 and 5 for Case 1 (when domain level auxiliary information was available at the PSU level) and 6,7 and 8 for case 2 (domain specific auxiliary information is

available at the SSU level only for the selected PSUs) explains that, all the proposed calibrated regression type estimator of the domain total under two stage sampling design, is performing at par with the well established HT estimator from the criteria of %RB when selection of sample at various stages were done using SRSWOR sampling design. With SRSWOR at various stages of selection under two stage sampling design at the domain level, the HT is an design based unbiased and consistent estimator and for all possible sample sizes drawn from the population for a respective sample sizes of PSUs and SSUs, it will return %RB as 0, but as we are unable to evaluate the estimator for all possible sample sizes and considered a sample of 10000 as a substitute for our limited simulation study, the %RB value of the HT estimator is nearing zero. Further, our proposed estimator under Case 1 (when domain level auxiliary information was available at the PSU level) and Case 2 (domain specific auxiliary information is available at the SSU level only for the selected PSUs), the %RB is turning out to be slightly more than the HT estimator as GREG estimator (Deville *et al.*, 1992) is usually a biased estimator with respect to the population parameter and it is different than the classical regression estimator of Hansen *et al.* (1953). GREG estimators are generally design consistent estimators and the fact was depicted with the results.

From the Tables 3, 4 and 5 for Case 1 (when domain level auxiliary information was available at the PSU level) and 6,7 and 8 for case 2 (domain specific auxiliary information is available at the SSU level only for the selected PSUs), we can observe that, the proposed domain calibration regression type estimators, under two stage sampling design, were performing better than the HT estimator in all domains and for all combinations of sample sizes drawn in each domain for the criteria of %RRMSE. There was significant improvement in the efficiency of the estimator w.r.t. the HT estimator across all sample sizes. Further, there is significant improvement in the %RRMSE of the proposed estimator when the PSU level sample size is fixed and SSU sample sizes increase and when SSU level sample size is kept fixed and PSU

level sample size increases. Hence, it can be concluded that, the proposed domain calibration estimators at both PSU and SSU level are consistent and efficient estimators of domain total under two stage sampling design with respect to the HT estimator.

CHAPTER 4

CONCLUSIONS AND FUTURE RESEARCH

4.1 Introduction

This Chapter presents a summary of main findings and some concluding remarks from the research carried out in this project. In the next section, we set out the findings from the project and in Section 4.3, we provide some future research areas which need further attention.

4.2 Major Findings

Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources, see for example, Deville and Sarndal (1992) and other reference therein for an overview on the topic. Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario, i.e. In India where Land records/khasra registers are available. Now total number of villages (clusters) in each Tehsil (stratum) is known but the total number of villages under the constituent crop (Rice, wheat etc.) in the mixture i.e. number of villages having the crop as Pure Stand, mixture-1, mixture-2... may not be available. Further, the number of selected villages within each tehsil is fixed, but the number of selected villages within each stratum under the crop as pure stand, mixture-1, mixture-2... is a random quantity. These different categories pure stand, mixture-1,

mixture-2 ... may be considered as Domains. Domain estimation is a crucial aspect of sample surveys that allows researchers to make accurate inferences about specific subgroups or domains within a population. In many cases, the primary goal of a survey is not only to estimate population parameters but also to provide reliable estimates for smaller groups or domains of interest. These domains could be defined based on demographic characteristics, geographical regions, or any other relevant criteria. Domain estimation involves the application of statistical techniques to estimate parameters specific to these subgroups. It allows researchers to gain insights into the variations and characteristics unique to each domain, enabling more targeted and informed decision-making.

Estevao and Sarndal (1999) first envisaged some important issues in the use of auxiliary information to produce design-based estimates for domains. They identified three types of design-based estimators and discussed two of these in detail. Hidiroglou and Patak (2001) in their paper entitled "Domain Estimation Using Linear Regression" introduced another concept of domain calibration estimation and its conditional properties of recognizable subsets (Rao, 1985) for various uni-stage sampling designs. Clement *et al.* (2014) developed an analytical approach for generating domain calibration estimator to enhance survey estimates. Hidiroglu *et al.* (2016) developed domain calibration estimators using direct and modified direct design weights under SRSWOR. It was observed that most of the work related to domain calibration estimation for the finite population parameters was mostly restricted to only uni-stage sampling designs. But the main aim of any developed methodology was to implement the same in improvement of the estimates obtained from real life surveys. Real life surveys are generally multistage in nature and methodologies based on uni-stage designs cannot be applied directly to these survey data. Further, ignoring the survey weights will lead to inconsistent estimates of the population or domain parameters (Wu *et al.*, 2020). Hence there is an urgent need for development of the domain calibration estimation under multi-stage sampling design. Further,

usually the most commonly used multistage design is two stage sampling design which was mostly used for various surveys conducted by the state and the central agencies of Government of India. Hence, the study “A Study on Domain Calibration Estimators under Two Stage Sampling Design” was proposed under the project.

However, most of the work related to calibration is restricted to only single stage or two phase sampling designs whereas in large scale surveys the most commonly used design is two or multi stage sampling design and hence we cannot use the developed calibration methodology for single stage or multiphase sampling design for multistage sampling design because it takes into account a complex set of auxiliary information. Hence there was a need to develop calibration estimators for multistage design in the presence of complex auxiliary information. In this project, we have considered the development of domain calibration estimators of the domain total under two stage sampling design when there was availability of auxiliary information both psu and ssu level. We have considered that the selection in each stage is independent and considered two situations of availability of auxiliary information for both at PSU and SSU level. The developed estimators were compared with the usual HT estimator of domains following Sarndal *et al.* (1992) empirically and found that all the estimators are performing at par with the HT estimator with respect to the criteria %RB and performing better than the HT estimator with respect to the criteria %RRMSE for different combinations of sample sizes of both PSUs and SSUs drawn out of each domain. To summarize, overall all the proposed domain calibration estimators were found to be better than the already available estimators of the population total under two stage sampling design for all the situations of availability of auxiliary informations for both PSU and SSU level.

4.3 Further Research Areas

There is a need for evaluation of the estimators based on the true estimator of variance and for that cause a suitable variance estimation method can be applied to the estimator to find out the results. Hence, for future research, variance estimation of the proposed estimators using suitable re-sampling techniques can be envisaged. Further, development of design based estimates at Small domain level or Small area level can also be considered using the proposed estimators to modify the existing synthetic and composite estimators under two stage sampling design scenario.

विशिष्ट सारांश

देविल और सरंडाल (1992) द्वारा प्रस्तावित कैलिब्रेशन दृष्टिकोण सर्वेक्षण अनुमान में सहायक जानकारी का कुशल उपयोग करने के लिए एक और तकनीकों में से एक है। जनसंख्या पैरामीटरों के आलावा, उप-जनसंख्याओं या डोमेन थे जिनके लिए अब उपयुक्त अनुमानों की आवश्यकता थी। इस अध्ययन में, डोमेन कुल के लिए एक डोमेन कैलिब्रेशन अनुमानकर्ता को दो मंजिली नमूनाकरण डिज़ाइन के तहत विकसित किया गया था जिसमें चयन के PSU और SSU स्तर पर सहायक जानकारी की उपलब्धता की मानना की गई। अनुमानकर्ताओं की विचलन और उसके संबंधित विचलन अनुमानकर्ताओं की भी पाई गई। प्रस्तावित अनुमानकर्ताओं की दो मंजिली नमूनाकरण डिज़ाइन के तहत सीमित सिमुलेशन अध्ययनों के माध्यम से भी सत्यापित किए गए थे जिनमें R सॉफ्टवेयर में एक कृत्रिम जनसंख्या को उत्पन्न किया गया था। प्रत्येक डोमेन से PSU और SSU नमूना आकारों के विभिन्न संयोजनों को खींचने के लिए निर्णय लिए गए थे। सिमुलेशन अध्ययन के माध्यम से पाया गया कि दो मंजिली नमूनाकरण डिज़ाइन के तहत सभी प्रस्तावित कैलिब्रेशन अनुमानकर्ता दो मंजिली नमूनाकरण डिज़ाइन के होर्विट्ज-थॉम्पसन अनुमानकर्ता के साथ प्रतिशत संबंधित प्रतिकूलता के संदर्भ में समान स्तर पर और %RRMSE के मानदंड के संदर्भ में बेहतर है।

EXECUTIVE SUMMARY

The Calibration Approach proposed by Deville and Sarndal (1992) is one of the other techniques widely used for making efficient use of auxiliary information in survey estimation. Beyond the population parameters, there were sub-populations or domains for which estimates were needed to be generated now days. In this study, a domain calibration estimator for domain total were developed under two stage sampling design with assumptions of availability of auxiliary informations at both PSU and SSU level of selection. The variance of the estimators and the corresponding variance estimators were also found. Both the proposed estimators were verified through limited simulation studies by generating an artificial population in R software. Various combinations of PSU and SSU sample sizes were drawn from each of the domains to draw the conclusions. Through the simulation study, it was found that all the proposed calibration estimators under two stage sampling design were performing at par with the Horvitz-Thompson estimator of the domain total under two stage sampling design with respect to the criteria of percent relative bias and performing consistently better than the HT estimator for the criteria of percent relative root mean square error.

REFERENCES

- Aditya, K., Biswas, A., Gupta, A.K. and Chandra, H. (2017). District level crop yield estimation using calibration approach. *Current Science*, **112(9)**, 1927-1931.
- Aditya, K., Sud, U.C. and Chandra, H. (2014). *Some calibration estimators under two-stage sampling design*. Project report. ICAR-IASRI, New Delhi.
- Aditya, K., Sud, U.C. and Chandra, H. (2016). Calibration approach-based estimation of finite population total under two stage sampling. *Journal of the Indian Society of Agricultural Statistics*, **70(3)**, 219-226.
- Aditya, K., Chandra, H., Kumar, S., and Das, S. (2019). Higher order calibration estimator of finite population total under two stage sampling design when population level auxiliary information is available at unit level. *Journal of the Indian Society of Agricultural Statistics*, **73(2)**, 99-103.
- Aditya, K, Sud, UC and Chandra, H (2014). Estimation of domain mean using two stage sampling with sub-sampling of non-respondents. *Journal of the Indian Society of Agricultural Statistics*, **68(1)**, 39-54.
- Aditya, K., Sud, U.C. and Hukum Chandra (2012). Estimation of domain total for unknown domain size in the presence of nonresponse. *Statistics and Applications*, **10**, Nos.1 & 2, pp. 13-25.
- Alam, S., Singh, S. and Shabbir, J. (2020). Calibrated estimators using non-linear calibration constraints. *Journal of Statistical Computation and Simulation*, **90(3)**, 489-514.
- Alam, S., Singh, S., and Shabbir, J. (2023). Optimal calibrated weights while minimizing a variance function. *Communications in Statistics-Theory and Methods*, **52(5)**, 1634-1651.
- Biswas, A., Aditya, K. and Sud, U.C. (2016). *Calibration estimators under two stage sampling design when Study variable is inversely related to auxiliary variable*. Project Report, New Delhi Publication.
- Biswas, A., Aditya, K., Sud, U.C. and Basak, P. (2020). Product type calibration estimation of finite population total under two stage sampling. *Journal of the Indian Society of Agricultural Statistics*, **74(1)**, 23-32.

- Biswas Ankur, Aditya Kaustav, Sud U.C. and Basak Pradip (2023). Calibration Estimator in Two Stage Sampling Using Double Sampling Approach when Study Variable is Inversely Related to Auxiliary Variable. *Statistics and Applications*, **21**, No. 1, pp 11-22.
- Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63(3)**, 615-620.
- Clark R. G. and Chambers R. (2022). Adaptive calibration for prediction of finite population totals. *Centre for Statistical and Survey Methodology*-working paper 05-08, The University of Wollongong, Australia.
- Clement, E. P. and Enang, E. I. (2014). Multivariate calibration estimation for domain in stratified random sampling. *International Journal of Modern Mathematical Sciences*, **13(2)**, 187-197.
- Deming, W. E., and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11(4)**, 427-444.
- Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87(418)**, 376-382.
- Deville, J. C., Särndal, C. E. And Sautory O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, **88**, pp. 1013-1020
- Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, **25**, 43-56.
- Estevao, V. M., and Särndal, C. E. (1999). Use of auxiliary informations in design based estimation of domains. *Survey Methodology*, **25**, No. 2, pp. 213-221
- Estevao, V. M., and Särndal, C. E. (2003). A new perspective on calibration estimators. *JSM-Section on Survey Research Methods*, **13**, 46-56.
- Estevao, V. M., and Särndal, C. E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, **74(2)**, 127-147.
- Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41(236)**, 517-529.

- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol. 1, Wiley, New York.
- Hidiroglou, M.A (2016). A comparison of small area and Calibration estimation-via Simulation. *Statistics in Transition*, **17(1)**, 133-154.
- Hartley, H. O. (1959). *Analytic Studies of Survey data*. Iowa state University Press.
- Hidiroglou M and Patak Z (2001). Domain Estimation Using Linear Regression. *Proceedings of Annual Meeting of American Statistical Association*. August 5-9, 2001.
- 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, E.T. and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *In Proceedings of the social statistics section, American Statistical Association*, pp. 300-305.
- Jerzy Neyman (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**, No. 4, pp. 558-625
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32(2)**, 133-142.
- Koyuncu, N., and Kadilar, C. (2010). On improvement in estimating population mean in stratified random sampling. *Journal of Applied Statistics*, **37(6)**, 999-1013.
- Koyuncu, N. and Kadilar, C. (2017). Calibration weighting in stratified random sampling. *Communications in Statistics-Simulation and Computation*, **45(7)**, 2267-2275.
- Koyuncu, N. (2018). Calibration estimator of population mean under stratified ranked set sampling design. *Communications in Statistics - Theory and Methods*, **47**, 5845-5853.
- Montanari, G.E. and Ranalli, M.G. (2005) Nonparametric Model Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association*, **100**, 1429-1442.

- Lehtonen R., Sarndal C.E., Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, **29**, 33-44.
- Lehtonen R., Sarndal C.E., Veijanen A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, **7**, 649-673.
- Mourya, K. K., Sisodia, B. V. S. and Chandra, H. (2016). Calibration approach for estimating finite population parameters in two-stage sampling. *Journal of Statistical Theory and Practice*, **10(3)**, 550-562.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–606.
- Ozgul, N. (2018). New calibration estimator based on two auxiliary variables in stratified sampling. *Communications in Statistics-Theory and Methods*, **48(6)**, 1481-1492.
- Ozgul, N. (2020). New improved calibration estimator based on two auxiliary variables in stratified two-phase sampling. *Journal of Statistical Computation and Simulation*, **91(6)**, 1243-1256.
- Raman, R. K., Sud, U.C., Chandra, H. and Gupta, V.K. (2013). Calibration estimator of population total with sub-sampling of non-respondents. *Journal of the Indian Society of Agricultural Statistics*, **67(3)**, 329-337.
- Rao, D., Khan, M.G.M. and Khan, S. (2012). Mathematical programming on multivariate calibration estimation in stratified sampling. *World Academy of Science, Engineering and Technology*, **72**, 78-82.
- Särndal, C. E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, **67(3)**, 639-650.
- Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99–119.

- Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag.
- Singh, S., Horn S. and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, **24(1)**, 41-50.
- Singh, S., Horn, S., Choudhury, S. and Yu, F. (1999). Calibration of the estimators of variance. *Australian and New Zealand Journal of Statistics*, **41(2)**, 199–212.
- Singh, S. (2003). *Advanced sampling theory with applications*. Dordrecht: Kluwer Academic Publisher.
- Singh S (2004) Golden and silver jubilee year-2003 of the linear regression estimators. *Proceedings of the American statistical association*, survey method section, Toronto: American Statistical Association, 4382–4389.
- Singh, S. (2006). Calibrated empirical likelihood estimation using a displacement function: Sir R. A. Fisher's Honest Balance. Presented at INTERFACE 2006. Pasadena, CA, USA.
- Singh, S., Arnab, R. (2011). On calibration of design weights, *Metron*, **LXIX**, 185-205.
- Sud, U.C., Chandra, H. and Gupta, V.K. (2014). Calibration based product estimator in single- and two-phase sampling. *Journal of Statistical theory and Practice*, **8(1)**, 1-11.
- Sud, U.C., Chandra, H. and Gupta, V.K. (2014). Calibration approach-based regression type estimator for inverse relationship between study and auxiliary variable. *Journal of Statistical Theory and Practice*, **8(4)**, 707-721.
- Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press. (USA).
- Tripathi, T.P. (1988). Estimation for domains in sampling on two occasions. *Sankhya*, **50**, 103-110.
- Tracy, D.S., Singh, S. and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, **29**, 99–104.
- Wu, C. and Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.

Wu C. and Thompson M E (2020). *Sampling Theory and Practice*. ICSA book series in Statistics. Springer.

Yates, F. (1953). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London.

Yates, F and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, **9(B)**, 223-261.

INDIAN COUNCIL OF AGRICULTURAL RESEARCH
RESEARCH PROJECT PROFORMA FOR MONITORING ANNUAL PROGRESS
(RPP- II)

(Refer for Guidelines ANNEXURE-XI (E))

1. Institute Project Code: AGEDIASRISIL202100100172
2. Project Title: A Study on Domain Calibration Estimators under Two Stage Sampling Design
3. Reporting Period: 18th January, 2021- 31st March 2021
4. Project Duration: Date of Start – 18th January 2021 Likely Date of Completion – 17th September, 2023
5. Project Team (Name(s) and designation of PI, CC-PI and all project Co-PIs, (with time spent for the project) if any additions/deletions

S. No.	Name, designation and institute	Status in the project (PI/CC-PI/ Co-PI)	Time to be spent (%)	Work components assigned to individual scientist
1.	Kaustav Aditya, Sr. Scientist, ICAR-IASRI, New Delhi	PI	30	Derivation, Writing R code, Simulation, Report Writing
2.	Pankaj Das, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing
3.	Raju Kumar, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing

6. (a) Activities and outputs earmarked for the year (as per activities schedule given in RPP-I)

Objective wise	Activity	Month & Year		Output monitorable target(s)	% to be carried out in different years			% achieved as targeted
		Start	Completion		1	2	3	
Objective 1	1. Relevant review of literature	January 2021	March 2021	Review	100	-		100

- (b) If shortfall/addition, reasons for the same and how to catch up with the intended activities

7. Annual Progress Report (research results and achievements in bullets)

- Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources, see for example, Deville and Sarndal

(1992) and other reference therein for an overview on the topic. Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario, i.e. In India where Land records/khasra registers are available. Now total number of villages (clusters) in each Tehsil (stratum) is known but the total number of villages under the constituent crop (Rice, wheat etc.) in the mixture i.e. number of villages having the crop as Pure Stand, mixture-1, mixture-2... may not be available. Further, the number of selected villages within each tehsil is fixed, but the number of selected villages within each stratum under the crop as pure stand, mixture-1, mixture-2... is a random quantity. These different categories pure stand, mixture-1, mixture-2 ... may be considered as Domains. Domain estimation is a crucial aspect of sample surveys that allows researchers to make accurate inferences about specific subgroups or domains within a population. In many cases, the primary goal of a survey is not only to estimate population parameters but also to provide reliable estimates for smaller groups or domains of interest. These domains could be defined based on demographic characteristics, geographical regions, or any other relevant criteria. Domain estimation involves the application of statistical techniques to estimate parameters specific to these subgroups. It allows researchers to gain insights into the variations and characteristics unique to each domain, enabling more targeted and informed decision-making.

- In this context a domain calibration estimator under two stage sampling design is needed to be developed. Relevant review of literature for the above study was done during this duration.
 - a. Trainings/demonstrations organized
 - b. Training received
 - c. Any other relevant information

8. Constraints experienced, if any: NIL

9. Lessons Learnt

10. Evaluation

(a) Self evaluation of the project for the period under report by the PI with rating in the scale of 1 to 10

10

(b) Evaluation by PI on the contribution of the team in the project including self

S. No.	Name	Status in the project (PI/CC-PI/Co-PI)	Rating in the scale of 1 to 10
1	Kaustav Aditya	PI	10
2	Pankaj Das	Co-PI	8
3	Raju Kumar	Co-PI	8

1. Signature of PI, CC-PI(s), all Co-PIs

कास्तव

(Kaustav Aditya)

पंकज

(Pankaj Das)

Raju Kumar

(Raju Kumar)

2. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of Head of Division/Regional Center / Section

8

कास्तव पंकज
06/10/2023

3. Comments of IRC

4. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of JD (R)/ Director

INDIAN COUNCIL OF AGRICULTURAL RESEARCH

**RESEARCH PROJECT PROFORMA FOR MONITORING ANNUAL PROGRESS
(RPP- II)**

(Refer for Guidelines ANNEXURE-XI (E))

1. Institute Project Code: AGEDIASRISIL202100100172
2. Project Title: A Study on Domain Calibration Estimators under Two Stage Sampling Design
3. Reporting Period: 1st April, 2021- 31st March, 2022
4. Project Duration: Date of Start – 18th January 2021 Likely Date of Completion – 17th September, 2023
5. Project Team (Name(s) and designation of PI, CC-PI and all project Co-PIs, (with time spent for the project) if any additions/deletions

S. No.	Name, designation and institute	Status in the project (PI/CC-PI/ Co-PI)	Time to be spent (%)	Work components assigned to individual scientist
1.	Kaustav Aditya, Sr. Scientist, ICAR-IASRI, New Delhi	PI	30	Derivation, Writing R code, Simulation, Report Writing
2.	Pankaj Das, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing
3.	Raju Kumar, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing

6. (a) Activities and outputs earmarked for the year (as per activities schedule given in RPP-I)

Objective wise	Activity	Month & Year		Output monitorable target(s)	% to be carried out in different years			% achieved as targeted
		Start	Completion		1	2	3	
Objective 1	2. Development of Calibration estimators of Domain Total when auxiliary information is available at PSU level under each domain.	April 2021	July 2021	Developed methodology	100	-		100
Objective 2	3. Development of Calibration estimators of Domain Total when auxiliary information is available at SSU level under each domain.	August 2021	November 2021	Developed methodology	100	-		100

Objective 3	4. Development of Variance and Estimate of Variance of the proposed Estimator of the proposed estimator under objective 1	November 2021	March 2022	Developed methodology	50	50		100
----------------	---	---------------	------------	-----------------------	----	----	--	-----

(b) If shortfall/addition, reasons for the same and how to catch up with the intended activities

7. Annual Progress Report (research results and achievements in bullets)

- Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources, see for example, Deville and Sarndal (1992) and other reference therein for an overview on the topic. Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario, i.e. In India where Land records/khasra registers are available. Now total number of villages (clusters) in each Tehsil (stratum) is known but the total number of villages under the constituent crop (Rice, wheat etc.) in the mixture i.e. number of villages having the crop as Pure Stand, mixture-1, mixture-2... may not be available. Further, the number of selected villages within each tehsil is fixed, but the number of selected villages within each stratum under the crop as pure stand, mixture-1, mixture-2...is a random quantity. These different categories pure stand, mixture-1, mixture-2 ... may be considered as Domains. Domain estimation is a crucial aspect of sample surveys that allows researchers to make accurate inferences about specific subgroups or domains within a population. In many cases, the primary goal of a survey is not only to estimate population parameters but also to provide reliable estimates for smaller groups or domains of interest. These domains could be defined based on demographic characteristics, geographical regions, or any other relevant criteria. Domain estimation involves the application of statistical techniques to estimate parameters specific to these subgroups. It allows researchers to gain insights into the variations and characteristics unique to each domain, enabling more targeted and informed decision-making.
- Estevao and Sarndal (1999) first envisaged some important issues in the use of auxiliary information to produce design-based estimates for domains. They identified three types of design-based estimators and discussed two of these in detail. Hidiroglou and Patak (2001) in their paper entitled "Domain Estimation Using Linear Regression" introduced another concept of domain calibration estimation and its conditional properties of recognizable subsets (Rao, 1985) for various uni-stage sampling designs. Clement *et al.* (2014) developed an analytical approach for generating domain calibration estimator to enhance survey estimates. Hidiroglu *et al.* (2016) developed domain calibration estimators using direct and modified direct design weights under SRSWOR. It was observed that most of the work related to domain calibration estimation for the finite population parameters was mostly restricted to only uni-stage sampling designs. But the main aim of any developed methodology was to implement the same in improvement of the estimates obtained from real life surveys. Real life surveys are generally multistage in nature and methodologies based on uni-stage designs cannot be applied directly to these survey data. Further, ignoring the survey weights will lead to inconsistent estimates of the population or domain parameters (Wu *et al.*, 2020). Hence there is an urgent need for development of

the domain calibration estimation under multi-stage sampling design. Further, usually the most commonly used multistage design is two stage sampling design which was mostly used for various surveys conducted by the state and the central agencies of Government of India. Hence, the study "A Study on Domain Calibration Estimators under Two Stage Sampling Design" was proposed under the project.

- However, most of the work related to calibration is restricted to only single stage or two phase sampling designs whereas in large scale surveys the most commonly used design is two or multi stage sampling design and hence we cannot use the developed calibration methodology for single stage or multiphase sampling design for multistage sampling design because it takes into account a complex set of auxiliary information. Hence there was a need to develop calibration estimators for multistage design in the presence of complex auxiliary information.
- In this project, we have considered the development of domain calibration estimators of the domain total under two stage sampling design when there was availability of auxiliary information both psu and ssu level. We have considered that the selection in each stage is independent and considered two situations of availability of auxiliary information for both at PSU and SSU level.
- The variance and estimator of variance of the proposed estimator when domain level auxiliary information was available at the PSU level was also developed.

8. Output During Period Under Report

- a. Special attainments/innovations
- b. List of Publications (one copy each to be submitted with RPP-II)
 - i. Research papers: Nil
 - ii. Reports/Manuals
 - iii. Working and Concept Papers
 - iv. Popular articles: Published an abstract entitled Kaustav Aditya, Bharti and Raju Kumar (2023). Domain Calibration Estimators under Two Stage Sampling Design when Population Level Auxiliary Information is Available at Cluster Level. In the proceedings of the 25th International Conference of SSCA during 15-17th February 2023 at Jammu University, Jammu.
 - v. Books/Book Chapters
 - vi. Extension Bulletins
- c. Intellectual Property Generation
(Patents - filed/obtained; Copyrights- filed/obtained; Designs- filed/obtained; Registration details of variety/germplasm/accession if any)
- d. Details of technology developed
(Crop-based; Animal-based, including vaccines; Biological – biofertilizer, biopesticide, etc; IT based – database, software; Any other – please specify)
- e. In many medium to large scale surveys, it is very often the case that we do not have a sampling frame. In some cases, the population could be spread over a wide area entailing very high travel expenses for the personal interviewers and efficient supervision of the field work can be difficult. In these situations, we prefer to use multistage sampling designs. Many times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large. In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of finite

population total or mean or distribution function. The Calibration Approach (Deville and Sarndal, 1992) is one of the techniques widely used for making efficient use of auxiliary information in survey estimation by providing new set of weights by modifying the sampling design weights using auxiliary information. Now, to address the problem of domain estimation and to improve the domain specific estimators under two stage sampling design scenario, a domain calibration estimator of the domain total is developed under two stage sampling design when population level domain specific auxiliary information is available at cluster level as well as SSU level. The variance estimators of the proposed estimators were also developed.

- f. Trainings/demonstrations organized
- g. Training received
- h. Any other relevant information

9. Constraints experienced, if any: NIL

10. Lessons Learnt

11. Evaluation

(a) Self evaluation of the project for the period under report by the PI with rating in the scale of 1 to 10

10

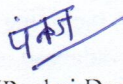
(b) Evaluation by PI on the contribution of the team in the project including self

S. No.	Name	Status in the project (PI/CC-PI/Co-PI)	Rating in the scale of 1 to 10
1	Kaustav Aditya	PI	10
2	Pankaj Das	Co-PI	8
3	Raju Kumar	Co-PI	8

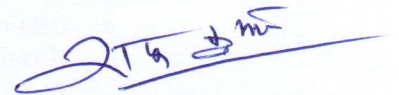
1. Signature of PI, CC-PI(s), all Co-PIs



(Kaustav Aditya)



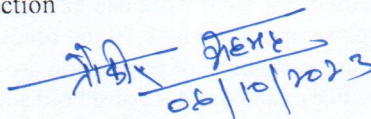
(Pankaj Das)



(Raju Kumar)

2. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of Head of Division/Regional Center / Section

8.5



3. Comments of IRC

4. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of JD (R)/ Director

INDIAN COUNCIL OF AGRICULTURAL RESEARCH
RESEARCH PROJECT PROFORMA FOR MONITORING ANNUAL PROGRESS
(RPP- II)

(Refer for Guidelines ANNEXURE-XI (E))

1. Institute Project Code: AGEDIASRISIL202100100172
2. Project Title: A Study on Domain Calibration Estimators under Two Stage Sampling Design
3. Reporting Period: 1st April, 2022- 31st March, 2023
4. Project Duration: Date of Start – 18th January 2021 Likely Date of Completion – 17th September, 2023
5. Project Team (Name(s) and designation of PI, CC-PI and all project Co-PIs, (with time spent for the project) if any additions/deletions

S. No.	Name, designation and institute	Status in the project (PI/CC-PI/ Co-PI)	Time to be spent (%)	Work components assigned to individual scientist
1.	Kaustav Aditya, Sr. Scientist, ICAR-IASRI, New Delhi	PI	30	Derivation, Writing R code, Simulation, Report Writing
2.	Pankaj Das, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing
3.	Raju Kumar, Scientist, ICAR-IASRI, New Delhi	Co-PI	15	Simulation, Report Writing

6. (a) Activities and outputs earmarked for the year (as per activities schedule given in RPP-I)

Objective 3	5. Development of Variance and Estimate of Variance of the proposed Estimator of the proposed estimator under objective 2	April 2022	July 2022	Developed methodology		100		KA
Objective 4	6. Development of R code for Empirical evaluation of the developed estimators under objective 1	August 2022	November 2022	R code for analysis/development of R programme for developed estimation methodology		100		KA
	7. Development of R code Empirical evaluation of the developed estimators under objective 2	December 2022	March 2023	R code for analysis/development of R programme for developed estimation methodology		50	50	KA

- (b) If shortfall/addition, reasons for the same and how to catch up with the intended activities

7. Annual Progress Report (research results and achievements in bullets)

- Calibration has established itself as an important methodological instrument in large scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources, see for example, Deville and Sarndal (1992) and other reference therein for an overview on the topic. Many a times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large or in case of estimation of crop area and yield at district level under mixed cropping scenario, i.e. In India where Land records/khasra registers are available. Now total number of villages (clusters) in each Tehsil (stratum) is known but the total number of villages under the constituent crop (Rice, wheat etc.) in the mixture i.e. number of villages having the crop as Pure Stand, mixture-1, mixture-2... may not be available. Further, the number of selected villages within each tehsil is fixed, but the number of selected villages within each stratum under the crop as pure stand, mixture-1, mixture-2... is a random quantity. These different categories pure stand, mixture-1, mixture-2 ... may be considered as Domains. Domain estimation is a crucial aspect of sample surveys that allows researchers to make accurate inferences about specific subgroups or domains within a population. In many cases, the primary goal of a survey is not only to estimate population parameters but also to provide reliable estimates for smaller groups or domains of interest. These domains could be defined based on demographic characteristics, geographical regions, or any other relevant criteria. Domain estimation involves the application of statistical techniques to estimate parameters specific to these subgroups. It allows researchers to gain insights into the variations and characteristics unique to each domain, enabling more targeted and informed decision-making.
- Estevao and Sarndal (1999) first envisaged some important issues in the use of auxiliary information to produce design-based estimates for domains. They identified three types of design-based estimators and discussed two of these in detail. Hidirolou and Patak (2001) in their paper entitled "Domain Estimation Using Linear Regression" introduced another concept of domain calibration estimation and its conditional properties of recognizable subsets (Rao, 1985) for various uni-stage sampling designs. Clement *et al.* (2014) developed an analytical approach for generating domain calibration estimator to enhance survey estimates. Hidirolu *et al.* (2016) developed domain calibration estimators using direct and modified direct design weights under SRSWOR. It was observed that most of the work related to domain calibration estimation for the finite population parameters was mostly restricted to only uni-stage sampling designs. But the main aim of any developed methodology was to implement the same in improvement of the estimates obtained from real life surveys. Real life surveys are generally multistage in nature and methodologies based on uni-stage designs cannot be applied directly to these survey data. Further, ignoring the survey weights will lead to inconsistent estimates of the population or domain parameters (Wu *et al.*, 2020). Hence there is an urgent need for development of the domain calibration estimation under multi-stage sampling design. Further, usually the most commonly used multistage design is two stage sampling design which was mostly used for various surveys conducted by the state and the central agencies of Government of India. Hence, the study "A Study on Domain Calibration Estimators under Two Stage Sampling Design" was proposed under the project.
- However, most of the work related to calibration is restricted to only single stage or two phase sampling designs whereas in large scale surveys the most commonly used design is two or multi stage sampling design and hence we cannot use the developed calibration methodology for single stage or multiphase sampling design for multistage sampling design because it takes into account a complex set of auxiliary information. Hence there was a need to develop calibration estimators for multistage design in the presence of complex auxiliary information.

- In this project, we have considered the development of domain calibration estimators of the domain total under two stage sampling design when there was availability of auxiliary information both psu and ssu level. We have considered that the selection in each stage is independent and considered two situations of availability of auxiliary information for both at PSU and SSU level.
- During this duration, the variance and estimator of variance of the proposed estimator when domain level auxiliary information was available at the SSU level was also developed.
- R codes for simulation study for both the developed estimators were done.

8. Output During Period Under Report

- Special attainments/innovations
- List of Publications (one copy each to be submitted with RPP-II)
 - Research papers: NIL
 - Reports/Manuals
 - Working and Concept Papers
 - Popular articles: Published an abstract entitled Kaustav Aditya, Bharti and Raju Kumar (2023). Domain Calibration Estimators under Two Stage Sampling Design when Population Level Auxiliary Information is Available at Cluster Level. In the proceedings of the 25th International Conference of SSCA during 15-17th February 2023 at Jammu University, Jammu.
 - Books/Book Chapters
 - Extension Bulletins
- Intellectual Property Generation
(Patents - filed/obtained; Copyrights- filed/obtained; Designs- filed/obtained; Registration details of variety/germplasm/accession if any)
- Details of technology developed
(Crop-based; Animal-based, including vaccines; Biological – biofertilizer, biopesticide, etc; IT based – database, software; Any other – please specify)
- In many medium to large scale surveys, it is very often the case that we do not have a sampling frame. In some cases, the population could be spread over a wide area entailing very high travel expenses for the personal interviewers and efficient supervision of the field work can be difficult. In these situations, we prefer to use multistage sampling designs. Many times, besides the overall estimates, the estimates for different subgroups of population are also required (Hartley, 1959) called as domains. For example, in a household survey, the survey statistician may be asked to provide separate estimates for the different household types, like one member households, two member households, etc. or in Agricultural Census Surveys, separate estimates may be generated based on operational holding size groups like marginal, small, semi-medium, medium and large. In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of finite population total or mean or distribution function. The Calibration Approach (Deville and Sarndal, 1992) is one of the techniques widely used for making efficient use of auxiliary information in survey estimation by providing new set of weights by modifying the sampling design weights using auxiliary information. Now, to address the problem of domain estimation and to improve the domain specific estimators under two stage sampling design scenario, a domain calibration estimator of the domain total is developed under two stage sampling design when population level domain specific auxiliary information is available at cluster level as well as SSU level. The variance estimators of the proposed estimators were also developed. R code for simulation study of both the estimators were done.
- Trainings/demonstrations organized
- Training received
- Any other relevant information

9. Constraints experienced, if any: NIL

10. Lessons Learnt

11. Evaluation

(a) Self evaluation of the project for the period under report by the PI with rating in the scale of 1 to 10

10

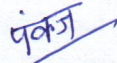
(b) Evaluation by PI on the contribution of the team in the project including self

S. No.	Name	Status in the project (PI/CC-PI/Co-PI)	Rating in the scale of 1 to 10
1	Kaustav Aditya	PI	10
2	Pankaj Das	Co-PI	8
3	Raju Kumar	Co-PI	8

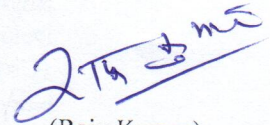
1. Signature of PI, CC-PI(s), all Co-PIs



(Kaustav Aditya)

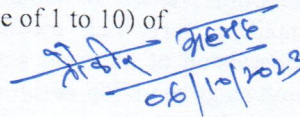


(Pankaj Das)



(Raju Kumar)

2. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of Head of Division/Regional Center / Section


06/10/2023

9

3. Comments of IRC

4. Signature (with specific comments on progress/achievements, shortfall and constraints along with rating of the project in the scale of 1 to 10) of JD (R)/ Director