CSIE Working Paper
#8

Second-Stage Stratification in India's Labour Force Surveys: A Simulation Study

Pratyush Priyadarshi, Anand Shrivastava and Amit Basole

June 2026

Centre for the study of the Indian Economy (CSIE)
Azim Premji University

Second-Stage Stratification in India's Labour Force Surveys: A Simulation Study*

Pratyush Priyadarshi,[†] Anand Shrivastava,[‡] and Amit Basole [§]

June 23, 2026

Abstract

In this paper, we examine the utility of second-stage stratification within the first-stage units (FSUs) of large-scale sample surveys, focusing on India's official labour force surveys, namely the NSSO Employment-Unemployment Survey 2011-12, the Labour Bureau Employment-Unemployment Survey 2013-14, and the Periodic Labour Force Survey 2022-23. These three surveys stratify households within the FSU in different ways, and we examine the implications of these choices for the statistical precision of the resulting estimates. Using Monte Carlo simulations on FSUs constructed from PLFS 2022-23 unit-level data, we estimate the variance of each design across three labour market outcomes: labour force participation rate (LFPR), unemployment rate (UR), and labour income. Our primary finding is that the household-level stratification designs employed in these surveys do not necessarily improve precision. For labour force participation and unemployment, simple random sampling outperforms stratification in most FSUs. In general, gains depend critically on the correlation between the stratification variable and the outcome of interest. We propose an alternative education-based scheme that combines lower secondary (10th pass) and graduate attainment. This design raises the median efficiency gain over the PLFS by about 14% for unemployment and 17% for income. We conclude with some reflections on the factors that are relevant to the choice of stratification design.

JEL Codes: C83, J21, J64, O15, C15

Keywords: Survey Sampling Methodology; Labour Statistics; Monte Carlo Methods

---

*We thank Anmol Somanchi and Prof G.C. Manna for helpful discussions. Views are authors' and do not necessarily represent those of their institutions.

†Independent researcher in Economics (Corresponding author), Bengaluru, India. Email: pratyush.p019@gmail.com

‡Azim Premji University, Bengaluru, India

§Azim Premji University, Bengaluru, India

1 Introduction

Large-scale, nationally representative sample surveys are a vital part of the economic statistics system in most countries. These surveys rely on a statistically grounded sample selection process whose objective is to produce estimates of economic variables that are unbiased and efficient.[1] One step in this process, used in many surveys, is to partition the ultimate-stage units (e.g. households) within a First Stage Unit (e.g. a village or an urban block) into mutually exclusive bins (strata) defined by pre-specified characteristics. Stratification is meant to create homogeneous sub-groups and thereby improve the efficiency of the survey estimates (Cochran, 1977, p. 89). In principle, then, the variance of a stratified estimate should be lower than that under Simple Random Sampling (SRS). It is worth noting that this step pertains to an estimate's efficiency, in the sense defined above, and not its unbiasedness.

In India, the surveys carried out by the National Statistical Office (NSO, formerly the National Sample Survey Organisation or NSSO) have been the mainstay of information on various aspects of development (such as consumption, employment, health, education, informal sector conditions and living standards) for the past seven decades. These surveys are also considered the gold standard in terms of their multi-stage sampling strategy. At each stage a set of units is selected by probability proportional to size without replacement (PPSWOR), with the unit changing from one stage to the next: districts, then towns or villages, and ultimately households (Figure 1). Within this system, the stratification step we focus on is called "Second Stage Stratification" (SSS) (MoSPI, 2001).

Step 1: Selection of geographic sampling frame
India is partitioned into NSS regions or districts, one of which is adopted as the primary sampling frame.

Step 2: First-stage stratification and selection of FSUs
Within the selected frame, urban areas are stratified into population size classes of towns and cities followed by PPS selection of urban FSUs; in rural areas, villages are selected using PPS.

Step 3: Second-stage stratification and household sampling
Within selected FSUs, households are listed, binned into a particular stratum using listing information, and samples are drawn independently from each stratum.

Figure 1: Three-step sampling design with second-stage stratification

The theory behind SSS is well established (discussed in detail in Sections 2 and 3). In brief, the stratification variable should be chosen so that the resulting groups are relatively homogeneous with respect to the outcome being measured (low intra-group variance) and distinct from one another (high inter-group variance) (Kish, 1965, p. 88). The aim is to make each stratum

[1] Unbiasedness of an estimator implies that the expected value of the estimator equals the true population value. The efficiency of an unbiased estimator refers to its having the lowest possible variance, implying that the estimate will be close to the true population value.

internally homogeneous by grouping the population into mutually exclusive bins. The more homogeneous a stratum, the better the estimate obtained from a unit drawn at random within it. The resulting improvement in precision is measured by the efficiency gain (Deaton, 2018, p. 15; Cochran, 1977, p. 99), expressed as the ratio of the variance under simple random sampling to that under stratified sampling.[2] A better sampling design implies a higher efficiency gain and a lower variance of the estimates.

However, this efficiency gain comes at a cost. SSS is only possible after a complete house listing exercise that collects information pertinent to the stratification exercise from each household in the FSU. Further, in practice, the choice of a stratification variable is complicated by other considerations such as the desire to ensure that households forming a tiny share of the population (e.g. very rich or very poor households, or certain minority groups) are represented in the sample (Cochran, 1977, p. 135). To make matters even more complex, a typical survey is expected to deliver information on several outcomes of interest, and a single stratification variable that meets the theoretical criterion for all of them may be difficult, if not impossible, to find (Cochran, 1977, p. 119; Kish, 1965, pp. 96–97). Whereas the NSSO provides detailed information on the sampling methods for all surveys, documentation on reasons for the adoption of a particular SSS strategy or reasons for change and their implications, is generally not available. This prevents researchers and others from arriving at an empirically grounded sense of the utility of the chosen SSS strategy.

Given that the SSS exercise is not costless, has practical difficulties, and that there does not exist much documentation justifying the choice of a strategy, it is worth asking if the strategies pursued in Indian sample surveys are capable of delivering efficiency gains as expected from theory. In this paper we use a simulation exercise to examine whether efficiency gains are likely to be achieved by the SSS methods used in India's national labour surveys. We evaluate the designs of three surveys, namely the NSS Employment-Unemployment Survey (EUS, in operation until 2011-12), the Labour Bureau Employment-Unemployment Survey (carried out between 2009-10 and 2015-16), and the more recent Periodic Labour Force Survey (PLFS, in operation since 2017-18). The SSS design has changed across these surveys. The NSS-EUS, for example, used Monthly Per Capita Consumption Expenditure (MPCE) as the stratification variable. The PLFS switched to educational attainment, forming strata from the number of household members who have completed ten years of schooling.[3]

Using Monte Carlo simulation on PLFS 2022-23 data, we investigate whether the estimates produced under a particular stratification design have lower variance than those under SRS. We

[2]The inverse of this term is called Design effect, which is also used in survey sampling literature. We discuss this further in Section 3.

[3]When the PLFS first started releasing numbers, there was controversy over whether the change in the SSS method rendered the EUS and PLFS non-comparable. We do not enter into this debate here, except to note that, with proper weighting, point estimates derived from any SSS method will be comparable. Whatever the variable chosen, the point of the exercise is to reduce the variance of the estimate by the process of stratification.

focus on three survey estimates: the labour force participation rate (LFPR), the unemployment rate (UR), and labour income. The evaluation has three steps. First, we construct a set of 100 FSUs by drawing from the PLFS data through random sampling proportional to household weights; these FSUs serve as the testing ground for the various SSS designs, and the way they are constructed stays the same across the labour surveys. Second, we apply each design, both existing and proposed, running 1,000 Monte Carlo iterations of every SSS strategy over all 100 FSUs to obtain a distribution of variance estimates for each FSU. Third, for each design and outcome we compute the efficiency gain as the ratio of the SRS variance to the stratified variance, and compare these gains across designs.

We find that the SSS methods in use do not deliver consistent efficiency gains. For most FSUs they fall short of a gain greater than one for LFPR and UR, the main exception being income under the NSS-EUS (MPCE) design. We then suggest some changes to the stratification scheme that could deliver more consistent gains.

The remainder of the paper is organised as follows. Section 2 reviews the history of sample surveys in India and describes the stratification designs used by the three labour surveys. Section 3 sets out the data and methods. Section 4 presents the results, and Sections 5 discusses the results and Section 6 concludes.

2 Background

2.1 A short review of sample surveys in India

PC Mahalanobis, a physicist turned statistician, is credited for the first successful implementation of representative sample survey in India using the modern survey sampling methodology. On his return to Kolkata (then Calcutta) from Cambridge in 1915, he became increasingly interested in practical problems of statistics, and formed an informal circle of people who were contributing to statistics under the Statistical Laboratory(Rudra, 1996). In 1937, Mahalanobis was approached by the Jute Committee formed by the Bengal Government to provide survey-based estimates for the jute crop, having already established a reputation through his work in meteorology, anthropometry, and agricultural statistics at the Statistical Laboratory.

At first, his plan for exploratory analysis of sample survey was met with scepticism by the Jute Committee, to which Ronald Fisher, who was visiting India at the time, responded by recommending his method, assuring the release of funds for the next four years of operation. This led to the advancement in framing of land under cultivation and sampling unit optimisation for a representative method of selection. By 1941, Mahalanobis was able to compare the estimates of his sample survey with those obtained from the Bengal government's census of the jute crop conducted in the same year. The sample survey estimates turned out to be within

2.8% of the full enumeration estimates, achieved at almost one-fifteenth the cost, effectively dispelling the scepticism over survey sampling (Mahalanobis, 1944). The result presented a proof-of-concept of representative sampling over a large area, which set the stage for institutional adoption of survey sampling methods as a state policy for faster, cheaper and reliable statistics, both in India and globally.[4]

National Sample Survey (NSS) was established in 1950 by the independent government of India, under the guidance of Mahalanobis, with active support of first Prime Minister Jawaharlal Nehru, for the purpose of recording statistics on consumption, crop yield and acreage, small scale enterprises and handicrafts, etc., providing the much needed empirical framework in planning for a country with widespread poverty. The administrative task of sampling was taken up by the NSS, with the Indian Statistical Institute (ISI) in Kolkata, itself established by Mahalanobis in 1931 (NSSO, 2004). The scaffolding of the sampling method setup by Mahalanobis serves as a backbone for the household surveys conducted by the organization to this day (Manna, 2025). According to D.B. Lahiri, the three notable contributions from Mahalanobis to sample surveys were pilot surveys, concept of optimum survey design i.e., extraction of most statistically viable estimate at a fixed cost, and Interpenetrating Network of Subsamples (IPNS). A prominent method developed by Mahalanobis for controlling the bias arising from a surveyor was IPNS, where different surveyors collect subsamples of the whole sample (in a given area), and the variance of the subsamples selected from each surveyor (for the same design) is used to identify anomalous behaviour. IPNS has since become an indispensable part of modern survey sampling methodology (Ghosh et al., 1999).

Although Mahalanobis was a towering figure in setting up the methodological and institutional framework, he was by no means alone. PV Sukhatme, a statistician by training who had studied and worked with the likes of Jerzy Neyman and Ronald Fisher, was responsible for setting up independent infrastructure for crop surveys and introduced practices that sometimes differed from those of Mahalanobis. Institutionally, he was the first Statistical Adviser (1940-1951) to the Imperial (later Indian) Council of Agricultural Research, and later Head of the Statistics Division of the UN Food and Agriculture Organization in Rome, contributing greatly to the promotion of statistical methods in estimation practice (Gore, 1997). Other significant contributions to the sampling theory and estimation methods came from scholars associated with or trained under Mahalanobis (at ISI, Kolkata), including R.R. Bahadur, S.N. Roy, R.C. Bose, D Basu, C.R. Rao, D.B. Lahiri, K.R. Nair, M.N. Murthy and other luminaries, most of whom were instrumental to the NSS sample survey setup right from the beginning (Ghosh et al., 1999).

The NSS-ISI arrangement of annual surveys for the Government of India stayed until 1970,

[4]Mahalanobis served the United Nations Statistical Commission in various capacities, including as Chair, contributing to the development of sample surveys worldwide from its very first session in 1946.

after which all administrative capabilities were put under National Sample Survey Organisation (NSSO), Ministry of Statistics and Programme Implementation (MOSPI). The new formation lent its hand on creating a new brand of survey sampling by having more specific surveys for a given operation, and maintaining a quinquennial calendar instead of annual. It also brought a quinquennial Urban Frame Survey to serve an updated sampling frame, specifically for the urban areas in between the period of decennial census, to reflect the changes in population and land utilisation from urbanisation (Manna, 2025; NSSO, 2004).

The other major change was in stratification at the FSU. The household samples of a given FSU were partitioned into agricultural and non-agricultural households prior to 1970 as a part of second stage stratification. Post 1970, the surveys began to use Monthly Per Capita Consumption Expenditure (MPCE), drawn from the listed households, as the stratification variable. The consumption and employment schedules were canvassed in the same NSS round, and MPCE from the consumption schedule was used to stratify the employment schedule. This was perhaps the first modification to the SSS since employment and labour-force questions were first added to the NSS questionnaire for urban households (ninth round, 1955-56). MPCE remained the norm for stratification for the rest of the century, providing a more nuanced basis for sample selection (Manna, 2025).

Over time changes in the economy demanded a higher frequency of employment estimates, and quinquennial statistics became inadequate to track changes. In the year 2009, the Labour Bureau, under the Ministry of Labour and Employment, started its own (short-lived) annual labour surveys. If we were to look through the formation of strata through the brief existence of these surveys from 2009 to 2016, it seems that the design of SSS has occupied the minds of survey designers, witnessing various changes, ranging from the change in the stratification variable from MPCE to the count of the working age members of the household, and the allocation count per FSU changing from 8 to 10 to 12 households (Labour Bureau, 2015). With the coming of the PLFS in 2017-18, the Labour Bureau surveys were discontinued and the SSS method changed once again- education level of household members became the SSS criterion (NSO, 2017). Considering the changing nature of strata formation over the years, it is important to look at the context under which they evolved, and the implications of these changes on statistical inference. With that in mind, we explore the various SSS methods adopted by labour surveys and quantify their effect over the change in variance of the estimates.

2.2 Labour surveys in India and SSS methods

From its inception in 1972-73 until 2011-12, the NSSO's quinquennial Employment-Unemployment Survey (EUS) was the principal source of labour statistics in India. Since 2017-18, the revamped EUS, now known as the Periodic Labour Force Survey (PLFS) has been responsible for labour statistics at an annual (and more recently a quarterly and monthly) frequency.[5] In the hiatus between the two, the Labour Bureau (an entity distinct from the NSSO, operating under the Ministry of Labour and Employment) conducted annual Employment-Unemployment surveys until 2016-17[6], although its operations had began in 2009. All three surveys have adopted different stratification techniques and we pick three survey rounds representing three separate designs, namely NSS-EUS 2011-12, LB-EUS 2013-14, and PLFS 2022-23.

The surveys follow a two-stage stratified sampling procedure. For urban areas, the first stage of stratification involves creating an exhaustive list of towns and cities in the aggregate unit (district or NSS region) and binning them into mutually exclusive bins created based on the size of the population. A sample of cities and towns is then chosen based on PPSWOR. The sample set of cities and towns represent their particular population sized area. To arrive at the sampling unit for selection of households in urban areas, latest update of the Urban Frame Survey is used, and the unit is termed as "urban block." For rural areas, the list of villages in the particular district or NSS region is divided into mutually exclusive bins and a set of villages are selected based on PPSWOR. The village or the urban block is the First Stage Unit (FSU) (Manna, 2025; Labour Bureau, 2015).

The process of construction of an FSU is standard across NSSO surveys and was followed by the Labour Bureau as well. The number of individuals for an FSU is capped at 1200, and the excess number of individuals are adjusted onto hamlet groups (hgs) for village and sub-blocks (sbs) for urban areas.[7] The second stage stratification within FSUs is designed to create more homogeneous strata. For this purpose, household-level variables used in stratification are collected from each household within the FSU during the household listing exercise. All households are divided into one or another mutually exclusive stratum and SRS is then used to select the sample households.

The three chosen surveys have distinct ways of stratifying households (Table 1). The stratification variable chosen in the NSS-EUS 2011-12 was MPCE, collected during the listing of households. The households lying in the top 10% percentile of the MPCE distribution were in the first stratum, the next 60% in the second, and the bottom 30% into the last. In the general case (without hamlet group of sub-group formation), two households are allocated to the first bin, four household to

[5] The report and unit-level data from the first PLFS round (July 2017-June 2018) was not released until May 2019.
https://indianexpress.com/article/business/govt-to-release-withheld-periodic-labour-force-survey-for-2017-18-5757388/

[6] The Labour Bureau EUS did not publish unit-level data during its existence, and the report for the survey from August 2016–March 2017 was withheld until December 2020.
https://www.business-standard.com/article/economy-policy/unemployment-peaked-to-4-year-high-during-demonetisation-govt-survey-119011001329_1.html

[7] With the formation of the new cluster of excess households, the sampling procedure gets modified, but we restrict our attention to the case where the population size does not exceed 1200 to demonstrate the design effect without loss of generality.

the second bin, and two household is allocated to the last bin (NSSO, 2011).

Table 1: Labour surveys and their stratification at FSU

Survey	Household (HH) Stratification Variable	STRATA			
		Stratum 01	Stratum 02	Stratum 03	Stratum 04
NSS-EUS	Monthly Per Capita Consumption Expenditure (MPCE)	Top 10% MPCE	Middle 60% MPCE	Bottom 30% MPCE	NA
LB-EUS	Working Age Members (WAM)	WAM = 1	WAM = 2	WAM = 3–4	WAM ≥ 5
PLFS	Lower Secondary / 10th completes (LS)	LS ≥ 3	LS = 2	LS = 1	LS = 0

The second survey in our consideration was conducted by the Labour Bureau with its fourth Employment-Unemployment survey (January 2014–July 2014). Although these surveys have now been discontinued, we consider this one for its distinct approach to stratification, using the count of household working age members (WAM) as the criterion. Table 1 gives the distribution of working age members used to classify the households within FSUs into a particular stratum (Labour Bureau, 2015).[8]

Table 2: Sample allocation for FSU strata

Survey	Stratum 01	Stratum 02	Stratum 03	Stratum 04
NSS-EUS	2	4	2	NA
LB-EUS	1	3	4	2
PLFS	2	2	2	2

The third survey is the ongoing annual Periodic Labour Force Survey started by the NSSO in 2017-18. This again marks a deviation in stratification strategy with education being used for the binning process. The stratification parameter uses the count of household members who have completed at least 10th standard (lower secondary education), with strata defined by whether the household has no such member, one, two, or three or more (Tables 1 and 2) (NSO, 2022).

3 Data and methods

We make use of unit-level data from PLFS 2022–2023 to construct First Stage Units (FSUs) derived from a typical household listing exercise, where households are classified into particular strata according to the stratification design of the respective survey. The stratification designs are then compared through their sampling distributions for conformity with the Central Limit Theorem, after which we demonstrate the utility of Second Stage Stratification (SSS) for survey design through an empirical exercise.

[8]There is a flaw in the allocation strategy. The first stratum has only one household whereas at least two are needed for the variance calculation within a particular stratum (Kish, 1965, p. 78)

Comparative analysis of surveys on the basis of their precision and variance can be traced to Jerzy Neyman's landmark 1934 paper(Neyman, 1934), where he demonstrates the effectiveness of stratified random sampling over purposive sampling. Similar variance-based comparisons of survey designs were later developed in Yates and Cochran's 1938(Yates and Cochran, 1938) work on the analysis of groups of experiments. Following through this history, we find the term "design effect" being popularised in the usage of standard sampling theory textbooks(Kish, 1965, p. 88), which is the ratio of variance of stratified sampling design over the simple random sampling. We define efficiency gain as inverse of this ratio, which is more demonstrative for the function of our analysis.

The expression for efficiency gain is given by:

$$\text{Gain}_{\text{strat}} = \frac{\text{Var}_{\text{SRS}}(\hat{\theta})}{\text{Var}_{\text{strat}}(\hat{\theta})}$$

where:

- $\hat{\theta}$ is the estimator of interest (e.g., mean, proportion),
- $\text{Var}_{\text{SRS}}(\hat{\theta})$ is the variance under simple random sampling,
- $\text{Var}_{\text{strat}}(\hat{\theta})$ is the variance under stratified sampling.

In other words, a ratio greater than one reflects a gain in precision of the stratified design over an unstratified one.

3.1 FSU construction

As mentioned earlier, the sampling for an FSU is standard across the labour surveys conducted by the NSSO and the brief period in which the Labour Bureau took over the exercise. This allows us to experiment with the design effect of stratification under equivalent conditions and draw conclusions on their compatibility. The task at hand then becomes to construct an FSU which allows us to experiment with all such design strategies.

To construct an FSU, we use existing sample data from PLFS 2022–2023, specifically the urban region, which is a total of roughly 45,000 households. We construct 100 variations of FSUs with a total of no more than 1,200 individuals in each FSU. The households sampled for the FSU are drawn from the PLFS data using probability proportional to size without replacement (PPSWOR), with the PLFS household weights as the measure of size. The constructed FSU successfully replicates the aggregate distribution of key variables in urban areas and is therefore appropriate for testing different SSS strategies. The FSUs thus created reproduce the necessary relations between age, gender, caste, employment characteristics, consumption, and income.

Table 3 reports the aggregate characteristics of ten of the 100 constructed FSUs, alongside the PLFS sample.

Table 3: Summary Statistics Across Samples (FSUs and PLFS)

	Gender (%)		**Caste (%)**				**Labour Market (%)**		**Household Variables (INR)**	
Sample	**Male**	**Female**	**ST**	**SC**	**OBC**	**Others**	**LFPR**	**UR**	**Consumption**	**Income**
FSU_1	52.00	48.00	2.45	13.48	45.03	39.04	40.08	7.59	12485	25571
FSU_2	50.37	49.63	2.02	16.12	44.20	37.66	38.76	6.87	11882	21462
FSU_3	50.16	49.84	1.64	11.64	39.33	47.40	39.07	7.13	12426	23538
FSU_4	53.87	46.13	1.82	17.36	41.26	39.55	37.62	5.83	12412	20837
FSU_5	46.36	53.64	1.73	17.98	28.85	51.45	40.87	6.91	13167	28108
FSU_6	48.38	51.62	1.95	18.43	36.94	42.68	38.44	7.42	12609	23272
FSU_7	51.12	48.88	2.08	15.74	40.91	41.27	39.91	6.75	12735	24156
FSU_8	49.94	50.06	1.88	14.97	38.56	44.59	40.11	7.01	12514	23648
FSU_9	50.83	49.17	2.11	16.22	39.74	41.93	39.48	6.52	12390	22954
FSU_10	49.57	50.43	1.91	15.63	40.18	42.28	39.72	6.98	12601	23841
PLFS	49.63	50.37	1.95	15.42	39.88	42.75	39.52	6.93	12542	23384

Note: The table reports aggregate characteristics for ten of the 100 constructed FSUs, with the PLFS sample shown for comparison. Consumption and income are monthly household totals in INR; the income outcome analysed in the paper is individual labour income (Section 3.2).

The FSUs generated from this exercise give us an anchor for measuring efficiency gains at the aggregate level, and if a sampling strategy is effective, it must at least demonstrate effectiveness at this aggregate level of FSU behaviour. This serves as a practical litmus test for design effectiveness, without requiring the exhaustive replication of FSUs across India, each of which has widely varying relationships among demographic characteristics.

3.2 Estimators and Outcome Definitions

The analysis focuses on three labour market outcomes: labour force participation rate (LFPR), unemployment rate (UR), and labour income. LFPR is defined as the proportion of individuals aged 15 years and above who are in the labour force under the Current Weekly Status (CWS) approach. UR is defined as the proportion of labour force participants who are unemployed. Labour income refers to monthly individual labour income, constructed as the sum of regular salaried income and self-employment income wherever reported. Throughout the Monte Carlo simulations, labour income is estimated only among employed individuals.

For each simulated sample, estimates are computed using weighted ratio estimators. Let w_i denote the sampling weight assigned to individual i. The LFPR estimator is given by

$$\widehat{LFPR} = \frac{\sum_{i\in S} w_i LF_i I(age_i \geq 15)}{\sum_{i\in S} w_i I(age_i \geq 15)}$$

where LF_i indicates whether individual i is in the labour force, and the sum runs over all sampled individuals.

The unemployment rate estimator is

$$\widehat{UR} = \frac{\sum_{i\in S_{LF}} w_i U_i}{\sum_{i\in S_{LF}} w_i}$$

where U_i indicates whether individual i is unemployed, and the sum runs over all sampled labour-force participants.

Labour income is estimated as

$$\widehat{\bar{Y}} = \frac{\sum_{i\in S_E} w_i Y_i}{\sum_{i\in S_E} w_i}$$

where Y_i is individual i's monthly labour income, and the sum runs over all sampled employed individuals.

For stratified designs, weights are calculated separately within each stratum as expansion factors,

$$w_i = \frac{N_h}{n_h}, \qquad i \in h,$$

where N_h denotes the total number of individuals in stratum h and n_h denotes the number of sampled individuals belonging to households selected from that stratum. These weights are subsequently used in the estimation of all labour market outcomes.

The summary statistics reported in Table 3 serve only to compare the aggregate characteristics of the constructed FSUs with the PLFS sample. Monthly consumption expenditure is measured at the household level using household consumption expenditure reported in the PLFS, while household income is obtained by summing labour income across all household members. From here on, all references to income refer to individual labour income.

3.3 Monte-Carlo simulation

Monte Carlo Simulation refers to a class of computational algorithms that uses repeated random sampling to approximate statistical properties of the design. This gives us the estimates of mean and variance for each employment outcome (LFPR, UR and income) for each stratification method. The random sample drawn during each iteration is independent of the previous iteration and hence provides an independent estimate with every iteration. If enough repeated samples are generated, we get a distribution of estimates reflective of the statistical property of the design.

We generate estimates from each stratification design by performing 1,000 sampling iterations on each of the 100 constructed FSUs. This provides the distribution of estimates, where the mean closely matches the sample estimate, as expected from the Central Limit Theorem (CLT). Figure 2 illustrates the sampling distributions of estimates for labour force participation rate (LFPR), unemployment rate (UR), and income, obtained from 1,000 iterations of four sampling designs: simple random sampling (SRS), Periodic Labour Force Survey (PLFS) stratification, Employment–Unemployment Survey (EUS) stratification, and a modified LB–EUS stratification. For each design, the histogram and overlaid kernel density curve show the empirical distribution of the estimates, while the red vertical line marks the true sample mean. Note that, for evaluating the SSS designs, we are concerned only with the variance of the estimate and not the mean.

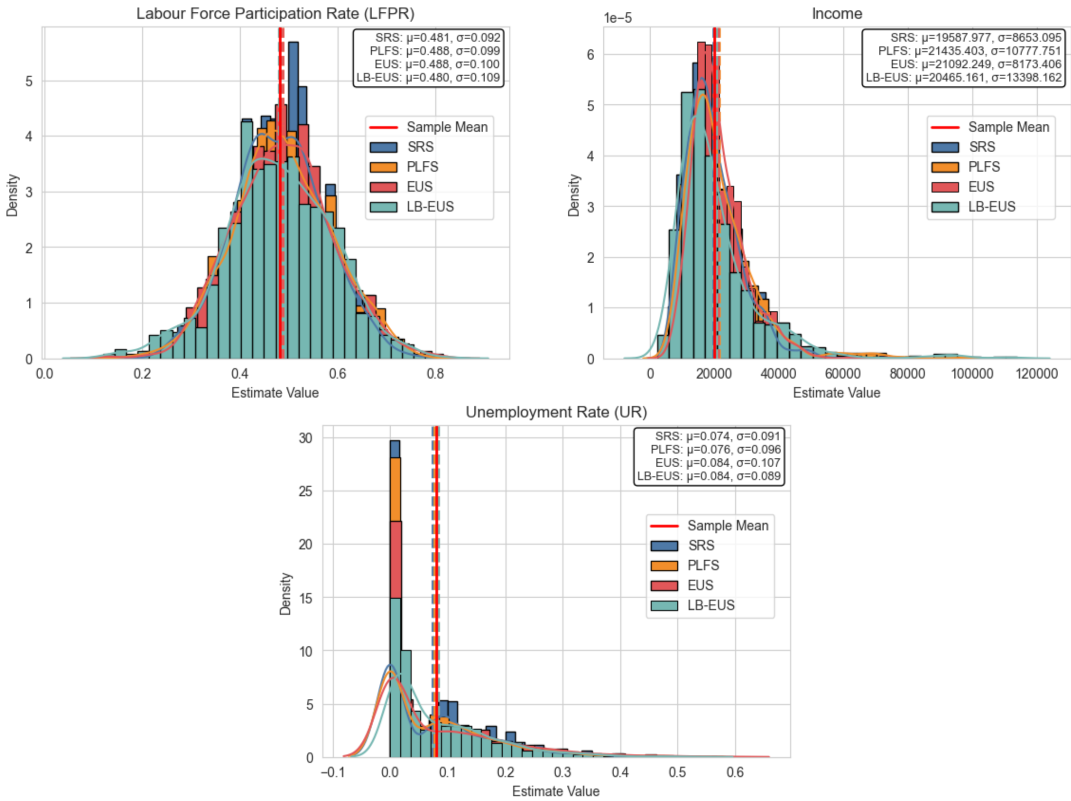

Figure 2: Central Limit theorem demonstration for LFPR, UR, Income

Consistent with the CLT, the distributions for LFPR and income are approximately symmetric and centred around the sample mean, with their spread determined by the design-specific standard deviations. The UR distribution shows a bimodal behaviour with zero as one of the modes due to lower chance of finding unemployed population from the random selection of households (in each iteration). The convergence of mean estimates across design illustrates that repeated sampling gives an unbiased estimate of the sample mean.

3.4 When is SSS useful? An illustration

Before proceeding to the efficiency gain calculations, we illustrate how stratification is, in principle, supposed to deliver more efficient estimates as compared to SRS. For the purpose of this illustration, we make use of FSU 1 from the sampled FSUs shown in Table 3 and LFPR is selected as outcome.

The purpose of stratification is to create a relatively homogeneous within-stratum population, and the better this is achieved, the more likely it is that efficiency gains are realised. For a given variance of the outcome in the whole population, the more homogeneous each stratum, the more separated the stratum means are from one another. We therefore measure the homogeneity achieved by stratification by the standard deviation of the stratum means.

To show how efficiency gains respond to homogeneity, we deliberately vary the homogeneity of FSU 1, starting from its PLFS strata. We progressively homogenise one stratum by transferring labour-force participants into it from the other strata, ignoring the variable on which the original strata were defined. Each transfer makes that stratum more internally homogeneous and pushes the stratum means further apart. At every step we run the Monte Carlo exercise and compute the efficiency gain as the ratio of the SRS variance to the stratified (PLFS) variance. Figure 3 shows the increase in the efficiency gain with the increase in the standard deviation of stratum means SD_{mean}.

$$\text{SD}_{\text{mean}} = \sqrt{\frac{1}{L}\sum_{h=1}^{L}(\mu_h - \bar{\mu})^2} \quad \text{where} \quad \bar{\mu} = \frac{1}{L}\sum_{h=1}^{L}\mu_h$$

Here, μ_h denotes the mean of stratum h, and L is the number of strata.

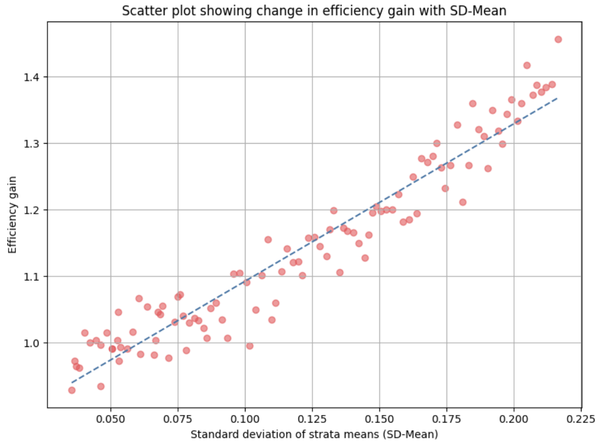

Figure 3: The variation in efficiency gain with the change in standard deviation of the strata means (SD_{mean})

The efficiency gain relative to simple random sampling is largely determined by the degree of separation between strata with respect to the variable of interest. When strata are more distinct (larger SD_{mean}), units within each stratum are more homogeneous, thereby reducing the variance of the stratified estimator. This reduction in estimator variance translates directly into higher efficiency gains.

The correlation between the stratification variable and the outcome plays a central role here. The illustration above manipulated homogeneity directly, but in a real design the only lever on homogeneity is the stratification variable. When that variable is strongly correlated with the outcome, the strata it forms are also well separated in the outcome, so the design can achieve a large SD_{mean} and hence substantial efficiency gains. This is the relationship we draw on in the results that follow: a design delivers gains for a given outcome only to the extent that its stratification variable is correlated with that outcome.

4 Results

4.1 Evaluating Stratification Effectiveness Across Constructed FSUs

The MCS process is replicated for the three stratification design principles (NSS-EUS, LB-EUS and PLFS) for three outcomes (LFPR, UR and income) and the efficiency gain is calculated along with SD_{mean}. Figure 4 shows the results.

Recall that the efficiency gain measure is obtained by dividing the variance under SRS by the variance under SSS. Thus a value above 1.0 reflects that the stratification design outperforms SRS in terms of precision. Starting with LFPR and UR across stratification designs, we see that the value is below 1.0 for the majority of First Stage Units (FSUs), suggesting that SSS does not perform better than SRS. The LB-EUS design shows a modest advantage in UR estimation but performs notably worse for LFPR, where gains consistently stay far below 1.0 for most of the FSUs. The NSS-EUS design performs especially poorly for LFPR, where the efficiency gain barely changes across FSUs, and for UR, where gains fall below one for most FSUs. PLFS retains some comparability with SRS, for both LFPR and UR, but performs suboptimally for most FSUs.

Figure 4: Scatter plot of efficiency gain from the constructed FSUs across the outcomes

The one notable strength of the NSS-EUS (MPCE-based) stratification design lies in income estimation, where its stratified output shows gains across nearly all FSUs. This aligns with the expectation that stratification can yield efficiency improvements when the stratification variable is correlated with the target variable, in this case consumption and income respectively. PLFS performs better than LB-EUS for income estimates, with a significant proportion of FSUs having a gain but the majority of FSUs still lie below one.

The foregoing results tell us that no single strategy is effective across the outcomes compared to SRS, and in several cases the stratification ends up with less precise estimates than SRS. Only the PLFS SSS strategy consistently shows some gains in the case of each outcome (though in all cases, more FSUs lie below 1 than above 1). So we take a closer look at the relation between educational attainment and labour market outcomes to redefine the stratification.

4.2 A new variation on the PLFS SSS strategy

Education levels are plausibly closely correlated with labour market outcomes, particularly the unemployment rate. To assess the relation between the education attainment and labour market outcomes, we plot the correlation between two binary indicators, namely whether the person has completed lower secondary school or 10th standard and whether the person has finished a diploma or degree, and the three labour market outcomes: LFPR, UR, and income. We also add a fourth outcome, a binary indicator for whether the individual has a regular wage job. We make use of the PLFS 2022-23 sample data at an NSS-region level of aggregation, the lowest level at which estimates are representative.

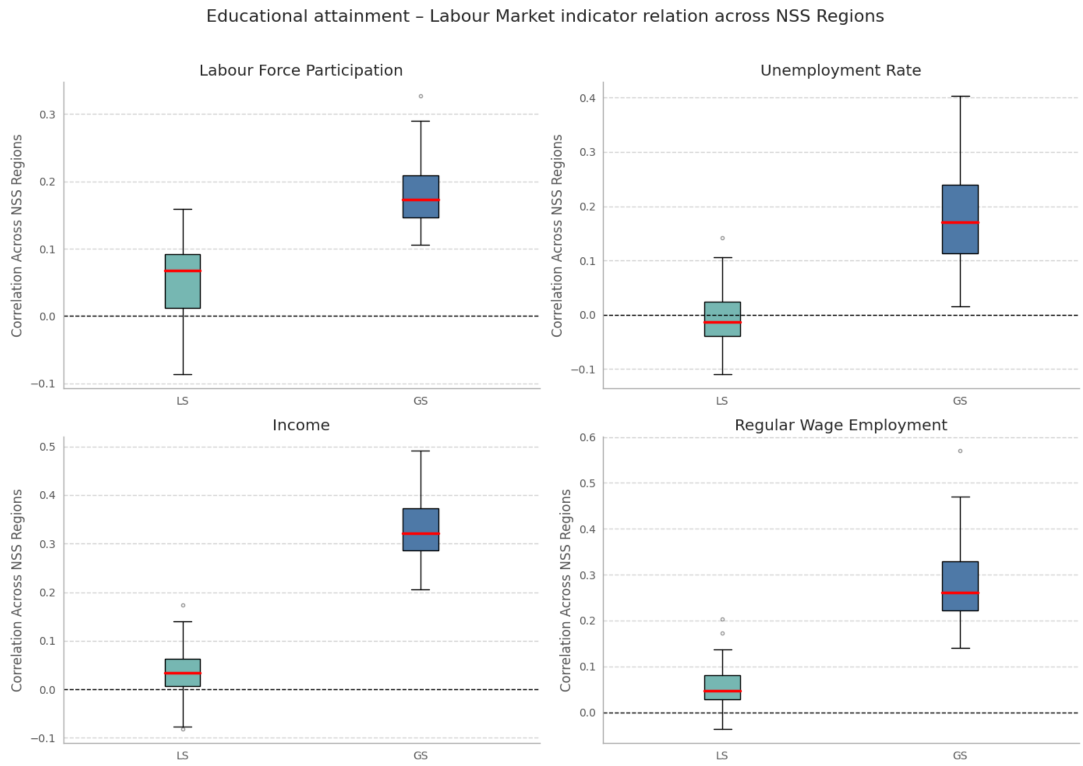

Figure 5: Correlation of two household education indicators (10th-pass and graduate) with labour-market outcomes across NSS regions.

Figure 5 shows how the two education indicators correlate with the labour-market outcomes across NSS regions, and the two attainment levels correlate quite differently. Graduate attainment is more strongly associated than lower-secondary attainment with being in the labour force, with being unemployed, and with higher income and holding a regular wage job. This suggests that graduate educational attainment may be useful when designing the stratification scheme.

As per the estimates from the PLFS 2022-2023, graduates make up 13.81% of the total population above the age of 24, with a concentration in the urban areas (27.81%), as compared to rural areas (8.21%). With the above discussed difference in labour outcomes between the

10th standard pass individuals versus graduates in mind, we define two stratification designs. One forms strata exclusively on the basis of the number of graduates in the household with a progressively increasing count of graduate members for the corresponding bins (GS method). The other design retains the PLFS scheme of stratification based on the 10th plus individuals for the bottom three strata but replaces the top stratum with graduates (LS-GS method). The allocation count for the two is the same as PLFS. The schema for the design and allocation is shown in Table 4.

Table 4: FSU stratification design using graduates

Survey Label	Household (HH) Stratification Variable	STRATA			
		Stratum 01	Stratum 02	Stratum 03	Stratum 04
GS	Graduate members in the households	GS ≥ 3	GS = 2	GS = 1	GS = 0
LS-GS	10th pass (LS) and graduate (GS) members	GS ≥ 1	GS = 0 & LS ≥ 2	GS = 0 & LS = 1	GS = 0 & LS = 0

Figure 6 shows the scatter plots of the efficiency gains with SD mean across FSUs. LS-GS performs relatively well on all three labour outcomes compared to GS with a significant number of FSUs crossing the 1.0 threshold, thereby surpassing the SRS outcome. The GS method is occasionally able to deliver exceptional efficiency gain on income and to an extent on UR, but for most FSUs, efficiency gain from GS falls short of crossing the threshold. Moreover, GS underperforms severely relative to LS-GS in estimating LFPR. This tells us that basing the stratification design strategy exclusively upon the number of graduates does not do much better than the original PLFS strategy. But perhaps a better design (at least for urban areas) would incorporate both graduate and 10th-plus educational attainment.

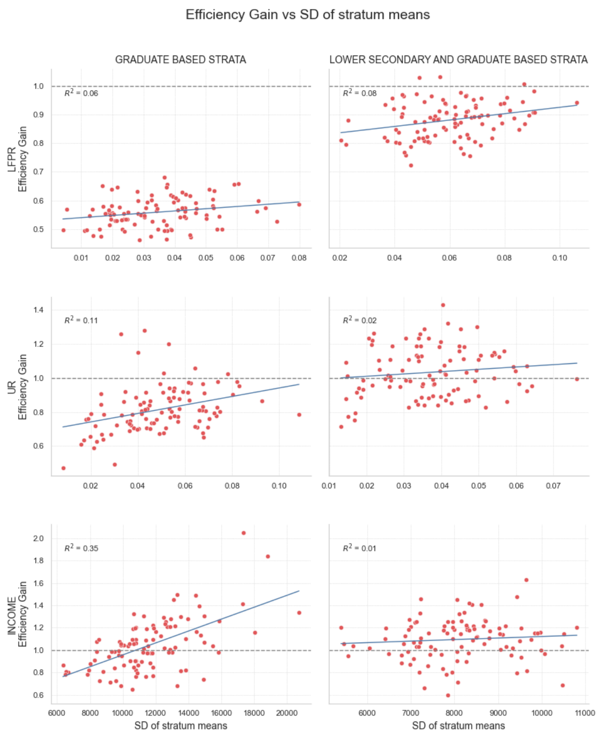

Figure 6: Scatter plot of efficiency gain with SDmean across FSUs for the graduate based stratification designs

Having 27.81% of graduates in the total urban population, along with a marked behavioural difference between the graduates and lower secondary completes, justifies accounting for graduates as a homogeneous stratum for a higher return on efficiency gains. As we observed in Section 3.4, the correlation between the design variable and the labour market indicator provides useful information for selecting the stratification design. However, we notice that the correlation in the combined LS-GS design is lower than that of only GS based strata. This primarily results from incorporating two distinct population groups, with marked differences in characteristics, into a unified schema, thereby weakening the direct relationship observed exclusively among graduates. Hence, although the correlation with labour market indicators turns out to be lower in the combined schema of the two, the design achieves a more representative stratification of the demography through the inclusion of both groups. This also implies that when designing stratification for sample surveys, it is important to incorporate multiple demographic characteristics to achieve higher efficiency gains, instead of relying exclusively on a unified characteristic.

4.3 Comparative analysis of efficiency gains

Finally, we look at the distribution of gains for both existing and suggested stratification designs using boxplots (Figure 7) to understand the relative differences in the gains distribution. We have three plots corresponding to the labour survey outcomes, LFPR, UR and income. In each plot we show the distribution of the efficiency gain parameter for each SSS design.

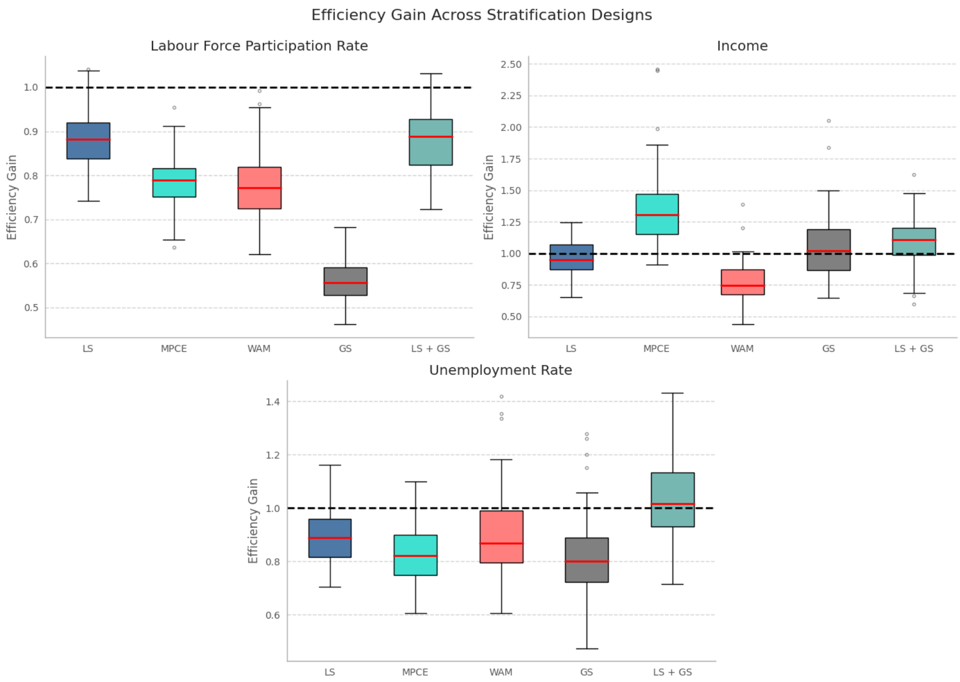

Figure 7: Box plot demonstrating the comparison of gains for different labour outcomes across stratification designs

Broadly, across SSS designs and outcomes, SRS remains more efficient as seen by the fact that the median remains below one for all cases except the new LS-GS design for UR and the MPCE and LS-GS designs for income.

For LFPR, PLFS and LS-GS perform better than others, although both designs underperform compared to SRS. The median efficiency gains for the unemployment rate outcome show relatively less variation across the design strategies compared to LFPR. The median of LS-GS crosses the 1.0 benchmark (although marginally), implying higher efficiency than SRS. Other designs fall short of crossing the 1.0 gain for the median, with the LS-GS median gain about 14% higher than that of PLFS, the next best design.

Finally, for the income estimates, MPCE-based SSS is able to deliver clear efficiency gains, with a median gain of 1.25 and maximum gains crossing 2.0. The PLFS (LS) and Graduate only (GS) designs show median gains below 1, while LB-EUS (WAM) design performs poorly across FSUs. The only other strata design maintaining a gain over 1.0 for the median of the outcome is LS-GS. The median gain for LS-GS outperforms PLFS by 17%.

Table 5: FSUs with efficiency gain greater than 1

Stratification Design	LFPR (%)	UR (%)	Income (%)
LS	6	14	36
MPCE	0	4	97
WAM	0	22	3
GS	0	7	53
LS-GS	3	54	73

Note: Designs are labelled by their stratification variable: LS = PLFS (10th-pass), MPCE = NSS-EUS, WAM = LB-EUS, GS = graduates only, LS-GS = combined 10th-pass and graduate. Cells give the percentage of the 100 constructed FSUs with an efficiency gain greater than one.

Overall, LS-GS offers a versatile design capable of delivering efficiency improvements across two of the three labour market outcomes. The marked difference is clearly highlighted in Table 5, which shows the proportion of FSUs with a gain greater than 1 under different stratification designs. In the case of UR, it produces a gain greater than 1 for 54% of the FSUs, far above the 14% of FSUs achieved by PLFS. Even for income estimates, it gives a proportion comparable to MPCE, with 73% of the FSUs recording a gain greater than 1, compared to 97% for MPCE, and more than double the proportion achieved by PLFS. No single design strategy offers universally optimal gains, which shows that stratification variables must be carefully aligned with the outcome of interest. Hence, survey designers must weigh the trade-off in efficiency across outcomes when selecting or adapting stratification strategies, particularly in the context of multivariable surveys.

5 Discussion

The foregoing analysis has four major lessons. Firstly, stratification does not automatically ensure an efficiency gain in the estimates. The three SSS strategies analysed did not show consistent gains compared to SRS for any of the three outcomes. In most cases, SRS outperformed stratified sampling. Consumption-based stratification (MPCE), as implemented in the NSS-EUS, showed notable efficiency gains for income estimation but performed poorly in estimating labour force participation and unemployment rates. The PLFS SSS method performed relatively better compared to those of the NSS-EUS and LB-EUS by holding consistent but with sub-optimal performance across the labour outcomes.

Secondly, it is evident that poorly designed stratification can hurt the efficiency of a survey estimate. The gain from a design depends on the correlation between the stratification variable and the outcome of interest. If the correlation is strong, a representative and efficient sampling strategy would make use of it to create intra-stratum homogeneity, and inter-strata heterogeneity. In the absence of knowledge of the relation between the demographic characteristic and the indicators, it is better to choose SRS on a complete listing of the households rather than a stratification strategy based on incomplete knowledge.

Thirdly, from our analysis of the relation between educational characteristic of the household to labour market outcomes, it follows that an education-based strategy stands as the best choice for efficient and representative sampling. We find that households with graduates need to be binned into a separate stratum as their labour market outcomes are substantially different. As PLFS bases its strategy exclusively on 10th-plus individuals, the stratification turns out to be inefficient on the labour market outcomes compared to the inclusion of graduates. The inclusion of graduates showed a characteristic improvement of median gain by 14% and 17% in the estimation of unemployment and income, respectively, compared to the currently adopted education-based stratification.

Fourthly, the gains from stratification depend on the demographic composition of each FSU, which varies across the country, so a stratification strategy should adapt to local composition rather than impose uniform strata everywhere.

In our analysis of the NSSO surveys, we have used efficiency gains to evaluate effectiveness of a particular stratification design. Here, the variance-based calculation is used for studying the stratification after it has been formed. However, the practice of modern survey methodology has been to use variance-based measures during the design stage itself in order to evaluate alternative sampling strategies before implementation. The European Social Survey, for example, requires the sample surveyors to evaluate the design effect measure of their stratification design before conducting the survey and to keep the measure below an upper bound (recall that, design effect is inverse of efficiency gain) (European Social Survey, 2022). Similarly, the Canadian Labour Force Survey uses coefficient of variation of a particular survey variable (such as unemployment rate) to decide upon the optimum design and sample allocation (Statistics Canada, 2017). Such practices ensure comparability both within and across surveys by constraining the variance of estimates within acceptable limits.

Underlying these practices is the use of prior demographic and survey information to guide the sampling design for a particular area unit. The modern sample survey methodology uses model assistance from the super-population structure, informed by previous census and sample surveys, to guide the effective design for the particular area unit, unlike the uniform design-based strategy used by the NSSO (Särndal, Swensson and Wretman, 1992; Rao, 2005). The uniform stratification design is relatively simple to implement but comes at the cost of an increase in overall variance of the estimates. The reason is that each of the FSUs distributed across the geographical boundaries of India has its own demographic traits, which are linked to survey outcomes (labour market outcomes in our case), and unless the sampling strategy makes use of it, a uniform stratification is likely to be ineffective in tapping into that relation.

We should also note that our preferred strategy too relies on a static distribution of household education, i.e., one stratum for graduates and three strata for distributing the 10th-complete households. This would often prove inadequate for a dynamically changing demographic composition across FSUs. For example, an FSU containing a disproportionately high number of graduates would have to be accommodated within a singular stratum, while an FSU with no graduates would face similar limitations. The census exercise in such instance would be a crucial link to create the relation between the survey outcomes and stratification variable, as we have seen in the case of Canada and Europe. This underscores the importance of a regular census exercise to ensure the strength of sample surveys.[9]

If census information becomes out of date, this makes the listing exercise, owing to its role in generating auxiliary information on the area, central to sample survey practice. In addition to this, the heterogeneity in the demographic characteristics of households within a particular village or urban unit observed in India makes listing an essential component. We suggest that in such circumstances, to replicate an optimally representative process of selection, the information collected during the listing of the household education should guide the design choices. The LS-GS survey methodology suggested earlier could take into account the percentage of households with graduates and lower secondary completes in a given area to arrive at strata formation with optimum design and allocation. This would enable the sampling design to respond to local demographic variation within FSUs by maintaining representativeness and statistical efficiency across heterogeneous survey units.

6 Conclusion

The NSO surveys continue to be critical to the generation of reliable, nationally representative information on various aspects of India's development performance. This study has placed the focus on a relatively under-emphasized aspect of sample survey design in India, the "second stage stratification" of households in a first stage unit (a village or an urban block in the case of NSSO surveys). Stratification may be adopted to achieve a variety of distinct goals. Ensuring representation of rare groups (such as the very rich or the very poor or minority groups) is one such goal. The improvement of the efficiency of sample estimates is another. We have shown, using a simulation exercise, that if the second goal is intended it is important to understand the relationship between the stratification variables and the outcome of interest.

We have argued that in multi-purpose surveys, where the relationship between easily obtainable information at the listing stage (which forms the basis for stratification)and several outcomes is hard to pin down, simple random sampling in FSUs may be more effective than a stratified sampling strategy that relies on a variable with a weak relationship to the outcome(s) of interest.

[9]Note that the population census scheduled to begin in 2021, has only recently begun its listing exercise in April 2026, and the enumeration will start in March 2027 (Registrar General and Census Commissioner of India, 2026).

Such a strategy increases the heterogeneity within the stratum, making the sample extracted from the population more variable and the estimate less efficient. By contrast, stratification on variables that are well-correlated with outcomes allows for more efficient sampling. We have proposed one such strategy based on educational characteristics of individuals in the household.

We hope that this study enables a more informed debate over the choice of household stratification strategies in multi-purpose sample surveys.

References

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

Deaton, A. (2018). *The Analysis of Household Surveys*. World Bank Group.

European Social Survey (2022). European social survey round 11 sampling guidelines: Principles and implementation. Technical report, European Social Survey.

Ghosh, J. K., P. Maiti, T. J. Rao, and B. K. Sinha (1999). Evolution of statistics in india. *International Statistical Review*.

Gore, A. P. (1997). P.v. sukhatme: A 'social' statistician. *Economic and Political Weekly*.

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

Labour Bureau (2015). Report on fourth annual employment & unemployment survey (2013–14). Technical report, Labour Bureau, Ministry of Labour & Employment, Government of India, Chandigarh.

Mahalanobis, P. C. (1944). On large scale sample surveys. *Philosophical Transactions of the Royal Society of London. Series B*.

Manna, G. C. (2025). *75 Years of the Indian National Sample Survey*. Springer.

MoSPI (2001). Concepts and definitions used in nss. Technical report, Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

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*.

NSO (2017). Note on sample design and estimation procedure. Technical report, Periodic Labour Force Survey 2017–18, National Statistical Office, MoSPI, Government of India.

NSO (2022). Note on sample design and estimation procedure. Technical report, Periodic Labour Force Survey 2022–23, National Statistical Office, MoSPI, Government of India.

NSSO (2004). Evolution of the sample design in the indian national sample survey from 1st to 55th round. Technical report, National Sample Survey Organisation, New Delhi.

NSSO (2011). Note on sample design and estimation procedure of nss 68th round (2011–12). Technical report, National Sample Survey Office, MoSPI, Government of India.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*. Statistics Canada.

Registrar General and Census Commissioner of India (2026). Press information bureau. Press Information Bureau, Government of India. 30 March 2026.

Rudra, A. (1996). *Prasanta Chandra Mahalanobis: A Biography*. Delhi: Oxford University Press.

Statistics Canada (2017). Methodology of the canadian labour force survey. Technical report, Statistics Canada.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.

Yates, F. and W. G. Cochran (1938). The analysis of groups of experiments. *Journal of Agricultural Science*.