This writing sample presents a comprehensive collection of my analyses and nearly all findings derived from the DHS 2019-21 India data. Though this snippet isn't the complete version, I'm happy to provide the full document if needed. To give you an overview: in this research, I confront the limitations of my earlier work published in the Journal of Progressive Research in Social Science by developing district-level mixed-effects spatial regression models to scrutinize educational attainment across Indian states, incorporating household-level random effects. This study rigorously examines demographic variables such as gender, caste, and religion, exploring how these factors influence inter-district spillover effects within various district clusters. We also delve into the impact of educational participation on wealth group transitions, employing a Markov chain analysis to mathematically derive the complexities of socio-economic mobility and the long-run possibilities for different demographic groups . The approach taken here offers a robust understanding of the intricate relationship between education, economic advancement and India’s caste/religion dynamic. I hope the data analysis in this sample may intrigue you to know more about my research. This work was conducted in R. If you're interested, I can provide the full version of this research. Your feedback would be invaluable.

Regards,
Arnab

2. Data Specification and Empirical Framework

2.1 NFHS Survey Overview

The National Family Health Survey (NFHS), executed on a quinquennial basis, has been the default dataset for a plethora of research endeavors over the last decades. It is conducted under the aegis of the Indian Ministry of Health and Family Welfare (MoHFW) and the National Sample Survey Office (NSSO). Due to its comprehensive scope—potentially the largest sample survey conducted in India—many significant research studies have utilized this data since the inception of the first NFH survey (NFHS-1) in 1992-93, despite critiques and limitations identified in previous iterations. For our study, we use the latest National Family Health Survey-5 (NFHS-5), the fifth installment, carried out in 2019-21. Focusing on the geographical distribution and causal channels of educational enrollment, NFHS-5 provides robust data on population, education, employment, and geography, and is possibly the only data source that covers all states, union territories (UTs), and 707 districts, meeting our needs. The survey was executed in two phases and amassed information from 636,699 households, 724,115 women, and 101,839 men. NFHS-5 builds on the foundations of previous surveys, starting from NFHS-1 in 1992-93, maintaining continuity in content and methodology while progressively expanding its scope. NFHS-4 (2015-16), with its enhanced sample size, provided district-level data and ensured comparability with earlier rounds. Similar to NFHS-4, NFHS-5 delivers district-level estimates for numerous crucial indicators, including the wealth index, which is extensively utilized in this study. It also introduces the new topic of preschool education, while continuing to cover incomplete education, access to schooling, reasons for drop-out, and occupation history. These elements enable a thorough analysis of educational participation and its socio-economic implications at both district and national levels, as well as the spatial spillover of educational barriers for our research . The survey design is as follows :

i. A uniform, stratified two-stage sample design was adopted, with districts stratified into urban and rural areas. Rural strata were further sub-stratified based on village population and the percentage of the population belonging to scheduled castes and scheduled tribes (SC/ST).

ii. Villages and Census Enumeration Blocks (CEBs) were chosen as Primary Sampling Units (PSUs), sorted by women's literacy rates in rural areas and by SC/ST population percentages. Rural villages were selected with probability proportional to size (PPS), and each rural stratum was divided into six approximately equal substrata. Urban CEBs were sorted similarly, and PSUs were selected using PPS systematic sampling.

iii. Each PSU or segment of a PSU had an estimated 100-150 households, with 22 households per cluster selected through systematic sampling.

iv. A total of 30,456 PSUs were selected across the country, with fieldwork completed in 30,198 PSUs. Four survey questionnaires—Household, Woman, Man, and Biomarker—were translated into 18 local languages and administered using Computer-Assisted Personal Interviewing (CAPI). Data was collected on household demographics, socio-economic characteristics, health insurance coverage, digital banking, Internet usage, land ownership, and mosquito net usage, among other topics. The survey also included extensive training and quality control measures.

v. Training of Trainers (ToT) workshops were conducted for field coordinators, who then trained fieldworkers at the state/UT level. Fieldwork was monitored by multiple levels of supervisors, including district coordinators, IIPS project officers, and senior staff from Field Agencies. Data quality was ensured through daily data transfers to IIPS, extensive data quality checks, and real-time feedback to field teams.

NFHS-5 achieved high response rates, with 98% of selected households successfully interviewed, 97% response rate for women, and 92% for men. The data collected provide valuable insights into India's health and family welfare landscape, assisting policymakers and program managers in setting benchmarks and assessing progress. This extensive and detailed survey is the largest sample survey conducted in India to date, offering critical data to inform public interventions and policy decisions.

2.2 Modelling District Level Educational Enrolment and Mapping the Distribution of Inter-Community Differences in Educational Participation

In this section, the district-level analysis aims to i) examine the variation in educational enrollment between and ii) within districts where the between district variation can have many causes starting from local governance, resource allocation, distance from the economic/political participation centers to historical context in term of both the initial condition and the series of events afterwards and enables us to visualize the map of this varying level of human development whereas the within district variation aims to identify which specific demographic factors have a stronger or weaker influence in educational participation deficit and understand in which parts of the country we see a higher concentration of inter community variation. This helps in understanding how different factors contribute variably to educational outcomes across regions and picture the landscape of incongruity to examine whether the inconsistency is barrier to participate (Borooah & Iyer, 2004) or lack of incentive to participate (Thorat & Attewell, 2007) or unequitable resource allocation (Jhingran & Sankar, 2009) at district level or a combination of them.

From a statistical perspective, using general decomposition techniques to address nested (or multilevel) country-level data may not yield satisfactory insights. This is not merely due to the sheer multitude of groups or the inconvenience of visualization and interpretability but also because nested data itself invokes two primary challenges. First, Spatial Heterogeneity, which pertains to spatial nonstationarity—this means that means, variances, and covariances vary across different regions. Second, there's Spatial Dependence, which is linked to spatial autocorrelation. For example, the relationship between educational attainment and gender might be strong in one district and weak in another, highlighting the inconsistency of underlying processes or relationships across the study area. This inconsistency is the problem of spatial nonstationarity. On the other hand, spatial autocorrelation measures the influence of one district on another. Models such as the Spatial Autoregressive Model (SAR) and the Spatial Durbin Model (SDM) are well-suited to study the second phenomenon. These models smooth out the autocorrelation effect and provide more accurate coefficients at a global level, offering a significant improvement over a global Ordinary Least Squares (OLS) model, if spatial autocorrelation is significant. However, these models just like OLS do not provide any district level measures and hence Ordinary Least Squares (OLS) or spatial global models are not suitable for our inquiry. Our goal is to identify spatial clusters, target districts needing intervention, and compare our

understanding of global participation deficits with a detailed district-level analysis of participation exclusion. Simply put, we are primarily interested in spatial heterogeneity. Therefore, it is more appropriate to refer to our approach as a "district-level" model rather than a "spatial" model since we are using districts to form clusters or groups.

Geographically Weighted Regression (GWR) is a widely utilized empirical instrument for analyses of this nature. GWR functions as a localized fitting method where regression coefficients are influenced by geographical location. This technique involves regressing each data point independently using a distance matrix, which determines the weights assigned to neighboring observations. The optimal bandwidth is selected before performing the regression, affecting how these weights are applied in the analysis. However, it is primarily suitable for conducting predictive analyses where the research question is, to some degree, pre-determined to elucidate the leveraging power of inter-boundary spillover effects in forecasting a particular dependent variable. Although GWR can address the type of between and within variability we aim to explore, its resource-intensive and time-consuming nature often leads to suboptimal data utilization. This is a significant trade-off that we seek to avoid in our study.

The districts and communities in India exhibit unique characteristics and idiosyncrasies that necessitate trial and error with various covariates and their functional forms. Additionally, each district-level sample possesses distinct limitations that are crucial for understanding important features. Conducting a thorough exploration of these features becomes particularly challenging when a GWR experiment, utilizing the Spgwr package in R, demands an estimated two weeks or more for a dataset comprising one hundred thousand points and five predictor variables (Harrish et al. 2010), Above all, the goal of our study is not predictive analysis.

To overcome the limitations of the methods we discussed so far, we adopt a more general approach that can include district level predictors (varying slope and varying intercepts), yet the individual region-specific estimates are adjusted borrowing strengths from the pooled information from all regions which is known as Compound Decision Problem. The classical Compound Decision Problem (or sometimes referred to as compound sampling model) is a concept where multiple decisions or estimations are required for a set of similar but not identical situations or unknown parameters θ's (such as unknown parameters for each district, in our case), but these θ's are realizations from the same unknown latent prior distribution.

To write it formally with mathematical notations :

$$Y_i \mid \theta_i \overset{indep}{\sim} p(\cdot \mid \theta_i) \;\; where\; \theta_i = X_i\beta_{ip} \tag{1}$$

Y is the variable of interest and for i^{th} district we refer that as Y_i. X_i represents matrix of observed $P\text{-}1$ auxiliary (independent) variables and intercept. And $\beta_i = (\beta_{i1}, \beta_{i2,} \dots \beta_{ip}) is$the i^{th} district specific parameters that we are trying to estimate. For ease of writing, we refer $X_i\beta_{ip}$ as θ_i. From the observed repeated realization of (Y_i, X_i) we estimate β_i for each district. In our problem, $p(\cdot \mid \theta_i)$ is likely Normally distributed. We will discuss the structure of it when we unravel the explicit regression formula that we propose.

We do not expect a global $\boldsymbol{\beta}$ because the parameters vary spatially. The 'compound' aspect of the problem lies in the fact that, while these β_i's are unique to each district, they are likely interrelated in

some manner. This problem is handled efficiently in Geographically Weighted Regression (GWR), where interconnectedness (in other words the neighborhood spillover) is captured using a weight matrix W_i (positive and symmetric), with each entry of the matrix is $w_{ij} = \exp\left(-\frac{d_{ij}^2}{2b^2}\right)$ or $w_{ij} = \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)$, where d_{ij} is the distance between the i^{th} district and the $j^{\text{th}} district\ and\ the\ estimated\ \beta_i = (X^T W_i X)^{-1} X^T W_i Y$. So, all $\beta_i\ depend$ on the observed X: The matrix of independent variables for all observations across all regions, and the observed Y: The vector of dependent variables for all observations across all regions . From the functional forms of weight matrix entries, we see that as the distance between i^{th} district and the $j^{\text{th}}\ district$ increases , the lesser β_i depend on a distant X_j as well as a distant Y_j . In our study we do not use GWR for its computational limitations, but we harness the notion of diminishing neighborhood effect to construct a Bayesian framework for our compound decision problem. Diverging from the conventional Hierarchical Bayesian or Empirical Bayesian models, we assert that the $\beta_i's$ do not emerge from a universal (country-level) latent distribution; rather, they manifest within spatial clusters each inscribed with a latent distribution of their own. Let us consider the identification of *S* spatial clusters, each inscribed with its own latent distribution G_s.

$$\theta_i \overset{iid}{\sim} G_k(.) \tag{2}$$

$k = 1, 2, \dots, S \quad and \quad i = 1, 2, \dots, n_k$ =number of districts in the k^{th} cluster

The reasoning behind not employing states (administrative level 1) as the second level of hierarchy lies in the possible misalignment of those boundaries with the geographical and historical contexts of Human development and Educational Enrollment. The reasoning behind not utilizing a general country-level latent distribution of hyperparameters lies in our methodological choice to refrain from pooling data across all districts. The distinct heterogeneity of each district precludes the assumption that region-specific parameters could be uniformly shrunk towards a country-level global mean, as it is done in a classical Empirical Bayes method.

In 1995, the 'Local Indicators of Spatial Association (LISA)' paper by Luc Anselin was published. This foundational work outlines the use of local indicators to analyze spatial association, providing a crucial tool for identifying clusters and spatial outliers in geographical data where Anselin defines a Local Indicators of Spatial Association is a function of Y_i and Y_{J_i} such that $Y_{J_i}$are the values observed in the neighborhood J_i of *i.*

Anselin proposes, a LISA measure as, $\mathrm{I_i} = \frac{(\mathrm{Y}_i - \overline{\mathrm{Y}})}{m_2} \sum_j w_{ij} \left(\mathrm{Y}_j - \overline{\mathrm{Y}}\right)$ (3)

Where m_2 is the variance of the variable of interest across all regions and $\overline{\mathrm{Y}}$ is the mean of the variable of interest across all regions (Anselin, 1995).

Using this metric we can identify four kinds of clustering :

High-High: Regions with high values surrounded by neighbors with high values, indicating 'hot spots'.
Low-Low: Regions with low values surrounded by neighbors with low values, indicating 'cold spots'.
High-Low: Regions with high values surrounded by neighbors with low values, indicating 'high outliers'.
Low-High: Regions with low values surrounded by neighbors with high values, indicating 'low outliers'.

The 'hot spots' and 'cold spots' are our spatial clusters. There can be clusters of either type, and it's important to note that there will be multiple 'hot spots' and 'cold spots.' All of these are spatial clusters, and we have S such clusters, as previously mentioned.

So, in our Bayesian framework if the i^{th} district is in k^{th} cluster then the θ_i is estimated from the Posterior distribution and Posterior mean are given by :

$$\text{Posterior Distribution : d}(\theta_i \mid Y_i) \propto P(Y_i \mid \theta_i) dG_k(\theta_i) \tag{4}$$

$$\text{Posterior Mean : } \mathrm{E}_{G_k}[\theta_i \mid Y_i] = \int \theta \; dG_k(\theta \mid Y_i) \tag{5}$$

Where the latent G's are generally determined from the prior belief. So, combining the (1) to (5) we get a likelihood of observed Y_i= $\{Y_{i1}, Y_{i2}, \ldots\ldots, Y_{in_i}\}$

$$Y_i \mid G_k \overset{iid}{\sim} f_{G_k}(.) \quad = \int p(Y_i|\theta) \; d \; G_k(\theta) \tag{6}$$

Now, we encounter two pressing dilemmas: a) the intricate task of discerning reasonable priors, and b) the nuanced endeavor of optimizing the likelihood function, culminating in the pursuit of optimal posterior inference. The most straightforward approach to address these challenges is employing the Hierarchical Bayes (HB) or Empirical Bayes (EB) solution, which we can implement without significant limitations.

In hierarchical Bayesian (HB) methods, regression coefficients are often given normal prior distributions. The introduction of Markov chain Monte Carlo (MCMC) techniques has resolved the need for explicit analytical solutions by using repetitive calculations to simulate samples from the posterior distribution. These simulated samples are then employed to derive important statistics, such as parameter estimates and confidence intervals. However, the challenge of determining parameters for normal priors still exists. Empirical Bayes (EB) methods address this by suggesting the use of observed data to estimate the prior distribution G.

So, under Normality assumption Y_i= $\{Y_{i1}, Y_{i2}, \ldots\ldots, Y_{in_i}\}$ follows a multivariate Normal Distribution:

$$Y_i = \{Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\} \sim N_{n_i}(\theta_i, \Sigma_i)$$

$$\text{And } \theta_i \mid \psi \sim N\left(\widehat{\theta_i}, \widehat{\psi_{n_i}}\right)$$

The James-Stein EB estimator for θ_i given by:

$$\widehat{\theta_i} = \lambda_i \overline{Y_i} + (1 - \lambda_i)\theta_0 \tag{7}$$

$$where\,, \lambda_i = \left(1 - \frac{(n_i - 2)\widehat{\sigma^2}}{\sum_{i=1}^{n_i} \overline{Y_i^2}}\right)$$

$$\overline{Y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad and, \; \widehat{\sigma^2} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y_i}\right)^2$$

$$\widehat{\psi_{n_i}} = \max\left\{0, \frac{1}{n_i}\sum_{i=1}^{n_i}\overline{Y_i^2} - \widehat{\sigma^2}\right\}$$

This is simple but powerful technique as James-Stein showed that his estimator dominates the maximum likelihood estimator (MLE) when $n_i \geq 3$ Specifically, the James-Stein estimator provides a better (i.e., lower mean squared error) estimate of the mean vector θ_i compared to the MLE, $\widehat{\theta_i} = \bar{Y}$ in higher dimensions. The key insight is that the MLE can be improved by "shrinking" the sample mean towards the overall mean, which reduces the overall estimation error. James-Stein's estimator leverages the borrowing effect by shrinking individual estimates towards the overall mean, thereby improving accuracy compared to the maximum likelihood estimator. Now we are moving on to implement the above framework for our problem.

I. Stage I Individual Level Model for a single district :

i. Model for district 'd' when we have adequate number of respondents from all demographic groups :

$$\text{Number of Years Education}_{ijd} = \beta_{0,d} + \sum_{k}\beta_{k,d}\cdot\left(\text{demographic factor}_k\right) + \sum_{p}\beta_{p,d}\cdot(\text{Gender}\times\text{demographic factor}) + \underbrace{\beta_{6,d}\cdot\exp\left(\frac{-1}{\text{age}_{ijd}^2}\right)}_{\text{represents the secondary schooling dropout}} + u_{j,d} + \epsilon_{ij,d} \quad \ldots\ldots \text{(i)}$$

Here $\text{demographic factor}_k \in \{\text{Muslim}, \text{SC/ST}, \text{Female}\}$; k = 1,2,3

; p =4,5

$\text{Number of Years Education}_{ijd}$ = Number of years of education for i^{th} individual from j^{th} household at the d^{th} district

$u_{j,d} \sim N(0, \sigma_u^2)$, Household level random effect for j^{th} Household at the d^{th} district.

$\epsilon_{ij,d} \sim N(0, \sigma^2)$, $k \in \{\text{Muslim}, \text{SC/ST}, \text{Female}\}$, $p \in \{\text{Muslim} * \text{female}, \text{SC/ST} * \text{female}\}$

- Interpretation of the β Estimates: While this model captures a significant percentage of variability, our aim is not predictive analysis. Rather, we seek to interrogate the distribution of the coefficients for Female (i.e., when the individual is female; female = 1), Muslim (i.e., when the individual is Muslim; Muslim = 1), and SC/ST (i.e., when the individual is SC/ST; SC/ST = 1).

But all districts may not have enough records to meaningfully execute the above model. Although the overall sample size for each district is sufficiently large for our study, there are many districts where we have very few

records for a particular demographic group (sometimes even less than 10 or 15 observations when the demography in that particular district is predominantly homogeneous) .

ii. Model for district 'd' when we do not have adequate number of Muslim respondents :

$$\text{Number of Years Education}_{ijd} = \beta_{0,d} + \beta_{female,d} \cdot \text{female} + \beta_{SC/ST,d} \cdot \text{SC/ST} + \beta_{female*SC/ST,d} * (\text{female} \times \text{SC/ST}) + \underbrace{\beta_{6,d} \cdot \exp\left(\frac{-1}{age_{ijd}^2}\right)}_{\text{represents the secondary schooling dropout}} + u_{j,d} + \epsilon_{ij,d} \quad \text{....(ii)}$$

iii. Model for district 'd' when it is a Muslim majority or an SC/ST majority district :

$$\text{Number of Years Education}_{ijd} = \beta_{0,d} + \beta_{female,d} \cdot \text{female} + \underbrace{\beta_{6,d} \cdot \exp\left(\frac{-1}{age_{ij,d}^2}\right)}_{\text{represents the secondary schooling dropout}} + u_{j,d} + \epsilon_{ij,d} \quad \text{....(iii)}$$

In our study we had 34 such districts where we have less than 100 general/upper caste respondents

II. Stage II estimates : Shrinking varying-intercept and varying-slope within spatial clusters

Following Stein's suggestion of $n_k > 3$ (n_k is number of districts/experiments in k^{th} cluster) using the NFHS data we identified 19 such clusters i.e. S= 19 . And for each of these 19 clusters we assume that the distribution of district level parameters follows latent distribution $G_1, G_2, \dots G_{19}$ and the latent distribution G formally plays the role of a prior distribution and that gives us our mixture distribution . For example, a cluster we identified in our study, encompassing 26 districts: Sheohar, Sitamarhi, Madhubani, Supaul, Araria, Kishanganj, Purnia, Katihar, Madhepura, Saharsa, and Darbhanga (in Bihar); Uttar Dinajpur, Dakshin Dinajpur, Maldah, Murshidabad, Birbhum, Nadia, Bankura, Paschim Medinipur, and Purba Barddhaman (in West Bengal); and Deoghar, Godda, Sahibganj, Pakur, Dumka, and Jamtara (in Jharkhand).

To note, many districts will not be part of any cluster if they do not have a positive LISA value (see equation 3) and that is acceptable.

Following the Stein estimator we stated in (7) , for districts at a cluster K (K = 1,2,…,S=19) we have the shrunk estimates as :

$$\widehat{\beta_{m,d}^{Stein}} = \lambda_d \mu_{\beta_m} + (1 - \lambda_d)\beta_{m,d} \quad \text{...(iv)}$$

Where, $\lambda_d = \max\left\{0, 1 - \frac{(n_K - 2)\widehat{\sigma^2}}{\sum_{i=1}^{n_K}(\beta_{m,i} - \mu_{\beta_m})^2}\right\}$ with $\mu_{\beta_m} = \frac{1}{n_K}\sum_{i=1}^{n_K}\beta_{m,i}$ and $\widehat{\sigma^2} = \frac{1}{n_K - 1}\sum_{i=1}^{n_K}(\beta_{m,i} - \mu_{\beta_m})^2$, is the shrinking factor.

In our exploration, we sought to build the mechanism that assesses both within and between variation, a task accomplished through the introduction of demographic covariates and the implementation of varying slopes and intercepts that are specific to districts and clusters. Our approach transcends the simplistic binary of urban-rural dummy variables, opting instead for a conceptualization that reflects the complex realities of spatial demographics. Rather than resorting to extensive regression tables laden with fixed effect components, we provide a concise and flexible mapping tool to represent the varying slopes and intercepts. This mapping tool does not merely depict data; it associates the spatial autocorrelation, weights, and spillover effects, within a spatially nonstationary structure. Eschewing the traditional Geographically Weighted Regression (GWR), we harness general regression tools to achieve this integration, a method that conserves time but also optimizes the use of information and obviates the need for specialized spatial econometric methods or expertise.

2.3 Effect of Education on Inter-Community Wealth and Social Class Transitions : Markov Chain Analysis

A Markov chain is a process that operates within a framework of defined "states," (in our study wealth category/social class) accompanied by a matrix that delineates the probabilities of transitions between these states over a fixed interval. At any given moment, the process resides within a singular state. Traditional Markov chain theory posits a singular subject navigating between states. However, in the realm of social mobility, the entire population is implicated, with each individual probabilistically shifting from one state to another. Within this context, an oft-implicit assumption emerges: Population homogeneity. (McFarland,1970) This presumes that all members of the population are subject to identical sets of transition probabilities, an assumption that can, at times, surreptitiously insinuate itself into the analysis. In our exploration, we attend to homogeneity within distinct demographic collectives: General/Upper Caste, SC/ST, and Muslim communities. Here, we discern unique patterns of social mobility, tracking the cumulative transitions from intermediate wealth to education, from education to occupation, and from occupation to final wealth. We calculate the wealth-to-education probabilities for individuals aged 18 to 25 years (to accurately account for the probability of attending tertiary education) , whereas the subsequent transitions from education to final wealth pertain to those aged 25 and older. Initially, we undertake this analysis on a national scale, revealing pronounced patterns, which we then scrutinize through a regional lens. This regional examination illuminates the relationship between spatial clusters of educational participation and regional effects on wealth mobility, identifying clusters of both low and high participation. These clusters, derived in the previous section through the identification of educational participation hot spots and cold spots, reveal a compelling association. This association manifests both broadly and within specific communities, clearly linking wealth mobility transitions to educational participation. We provide maps illustrating these spatial networks. The implications of this analysis are both urgent and optimistic, pointing

towards potential policy interventions. Before delving into the mathematical formulations, it is essential to clarify the metrics of education, wealth, and occupational hierarchy, as these quantifications are fundamental to calculating Markov transition probabilities.

2.3.1 Measurement Indicators : Education, Occupation & Wealth

In the NFHS 2019-21 data we get multiples variables to measure different aspects of human development . For the Markov Chain Analysis, we need to combine the variables into one single variable for each of the three aspects.

Education Related Variables : educational level, Highest year of education, whether completed the level of education, whether participated in a literacy program, whether attending school/college etc.

Wealth Related Variables : wealth index, wealth index factor score.

Occupation Related Variables : occupation, occupation(grouped), currently working/seasonally working etc.

Education:

Age Group	Combined Educational Participation Observation from NFHS 2019-21	Raw Score	E Participation Deficit Indicator	1-E
Children (5-12 years)	Not currently attending school	4	1	0
	Currently attending school or has attended school during the survey year	0	0	1
Adolescents (12-18 years)	Not attending school and has less than six years of schooling	4	1	0
	Not attending school but has six or more years of schooling, though has not completed secondary education	3	0.75	0.25
	Attending school, has six or more years of schooling, but still pursuing primary education	2	0.5	0.5
	Not attending school but has completed secondary education	1	0.25	0.75
	Attending school, has started or completed secondary education but not higher education, and has six or more years of schooling	0	0	1
Individuals (18 years and above)	Less than six years of schooling	4	1	0
	Six or more years of schooling, completed primary education but has not pursued secondary education, and is currently not out of school	3	0.75	0.25
	Six or more years of schooling, started secondary education but has not completed it, and is currently not attending school	2	0.5	0.5
	Six or more years of schooling, completed secondary education but not higher education, and is currently not attending school	1	0.25	0.75
	Attending school and has completed secondary education or higher, or not attending school but has pursued higher education	0	0	1

WEALTH:

COMBINED WEALTH INDEX FOR URBAN/RURAL = POOREST	**1**

COMBINED WEALTH INDEX FOR URBAN/RURAL = POOR	**2**
COMBINED WEALTH INDEX FOR URBAN/RURAL = MIDDLE	**3**
COMBINED WEALTH INDEX FOR URBAN/RURAL = RICH	**4**
COMBINED WEALTH INDEX FOR URBAN/RURAL = RICHEST	**5**

Occupation :

NFHS OCCUPATION GROUP	NFHS NUMERIC CODE	LABEL
PROFESSIONAL TECHNICAL MANAGERIAL	1	White Collar
CLERICAL	3	White Collar
SALES	4	White Collar
SERVICES HOUSEHOLD AND DOMESTIC	5	Blue Collar
AGRICULTURAL	6	Agricultural
SKILLED AND UNSKILLED MANUAL	7	Blue Collar

We performed robustness check with other available metrics indicators and the above indicators show good positive correlation.

2.3.1 Transition Matrix and other metrics for social mobility :

Markov chain Analysis is a statistical method used to model and analyze the mobility between different states (such as wealth levels and social classes) . It aids in understanding the long-term probability and the impact of factors like education on these transitions. By applying the mathematics of axiomatic probability, we can discern who is more likely to remain in the poorest state and who shows the most promise (or risk) of moving to adjacent states, whether upwards or downwards. We begin with a straightforward question: For an 18 to 25-year-old individual, what's the probability of starting in Wealth Category W_i and moving to Wealth Category W_j? If we had panel data tracking the same individuals over an extended period, this would be a simple matter of calculating the proportions within the aggregate sets of individuals with varying accomplishments. However, such extensive data is rarely available, particularly in a country like India, with its 1.3 billion people and immense socio-economic and geographic diversity. So, we break down the question into manageable parts: a) What is the probability of an individual from Wealth Group W_i achieving Education Level E_i b) What's the probability of an individual with Education Level E_i obtaining Occupation O_i? c) What's the probability of an individual in Occupation O_i accumulating Final Wealth W_j^f?

The calculation for each demographic community (homogeneity assumption) is as follows :

P_{w_j,w_i} = P (Final wealth = W_j^f |initial wealth = W_i) =
$\sum_{k=1}^{4} P\Big($ Final wealth = W_j^f $\big|$ Occupation = O_k $\Big) \cdot \sum_{i=1}^{3} P($ Occupation = O_k | E_i) ·
$P(E_i \mid W = W_i)$

For example,

For general/upper-caste group

a) P (E_i |W = 'middle') = % of 18-25 years old (excluding older age groups to reflect on the current state of education barrier) with education E_i and Wealth index = 'middle'.

From our data, the probabilities stand at:

P (E_i= 6 years of education or less | W = 'middle') = 0.0929476 …(1)

P (E_i= secondary or incomplete secondary |W = 'middle') = 0.4229990 …(2)
P (E_i = completed/pursuing higher |W = 'middle') = 0.4425967 …(3)

b) P ($Occupation = O_k|E_i$) for i = 1, 2, 3 can be calculated in the same manner from the % of 25+ years old (we change the subset to all employable age groups)

The combined probability P ($Occupation = O_k$|W = 'middle'), for general/upper-caste group will thus be :

$\sum_{i=1}^{3} P($ Occupation = O_k | E_i) · $P(E_i \mid W =$ 'middle') (*)

From our data, the probabilities stand at:

P (O_i= Not working | W = 'middle') = 0.1435975 …(5)

P (O_i= Agricultural labour |W = 'middle') = 0.2778188 …(6)

P (O_i = Domestic/skilled/unskilled manual labour |W = 'middle') = 0.2984799 …(7)

P (O_i = Professional/clerical/sales |W = 'middle') =0.2386471 …(8)

c) P (Final wealth = W_j^f |$Occupation = O_k$) for j = 1, 2, 3, 4, 5 can be calculated from the % of 18+ years old with occupation O_k and Household Wealth Index W_j^f:

$\sum_{k=1}^{4} P\Big($ Final wealth = W_j^f $\big|$ Occupation = O_k $\Big) \cdot P($ Occupation = O_k | $W =$ 'middle') (**)
= $\sum_{k=1}^{4} P\Big($ Final wealth = W_j^f $\big|$ Occupation = O_k $\Big) \cdot \sum_{i=1}^{3} P($ Occupation = O_k | E_i) ·
$P(E_i \mid W =$ 'middle')

by inserting (*) in (**)

= P (Final wealth = W_j^f |W = 'middle')

As a general expression :

P_{w_j,w_i} = P (Final wealth $= W_j^f$ |initial wealth = W_i) =
$\sum_{k=1}^{4} P\Big($ Final wealth $= W_j^f \mid$ Occupation $= O_k \Big) \cdot \sum_{i=1}^{3} P($ Occupation $= O_k \mid E_i) \cdot$
$P(E_i \mid W = W_i)$

We will have a set of 5x5 = 25 such probability expressions which can be set as a 5X5 Markov Transition Matrix.

Steady-State Distribution and Lerman-Yitzhaki Mobility Index : The steady-state distribution of a Markov chain represents the long-term behavior of the system. It shows the proportion of time that the system will spend in each state if it is observed over a long period. (In the context of our transition matrices, it indicates the long-term probabilities of individuals being in each wealth category unless there is any targeted intervention.) It is a probability distribution that remains unchanged as the system evolves over time. Mathematically, if $\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ is the steady-state distribution and P is the transition matrix, then :

$$\begin{aligned}
\pi_1 &= \pi_1 P_{w_1,w_1} + \pi_2 P_{w_1,w_2} + \pi_3 P_{w_1,w_3} + \pi_4 P_{w_1,w_4} + \pi_5 P_{w_1,w_5} \\
\pi_2 &= \pi_1 P_{w_2,w_1} + \pi_2 P_{w_2,w_2} + \pi_3 P_{w_2,w_3} + \pi_4 P_{w_2,w_4} + \pi_5 P_{w_2,w_5} \\
\pi_3 &= \pi_1 P_{w_3,w_1} + \pi_2 P_{w_3,w_2} + \pi_3 P_{w_3,w_3} + \pi_4 P_{w_3,w_4} + \pi_5 P_{w_3,w_5} \\
\pi_4 &= \pi_1 P_{w_4,w_1} + \pi_2 P_{w_4,w_2} + \pi_3 P_{w_4,w_3} + \pi_4 P_{w_4,w_4} + \pi_5 P_{w_4,w_5} \\
\pi_5 &= \pi_1 P_{w_5,w_1} + \pi_2 P_{w_5,w_2} + \pi_3 P_{w_5,w_3} + \pi_4 P_{w_5,w_4} + \pi_5 P_{w_5,w_5}
\end{aligned}$$

Where, $\pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1$ and π_i is the long-run probability that the system (the specific demographic group) be in state i.

Lerman-Yitzhaki Mobility Index : Finally, this simple measure will play a vital role in quantifying the net social mobility and the direction of mobility. The Lerman and Yitzhaki Mobility Index is another metric designed to measure the extent of mobility within a system, focusing on the changes in individuals' ranks within a distribution over time. Unlike other indices (e.g., Shorrocks or Bartholomew) it can be decomposed to show both upward and downward mobility. This makes it particularly useful for understanding the directional aspects of mobility along with the magnitude . The Lerman and Yitzhaki Mobility Index is based on the concept of rank changes. It quantifies (Downward Mobility) the extent to which individuals move to lower ranks and (Upward Mobility) the extent to which individuals move to higher ranks.

$$\text{Lerman-Yitzhaki Mobility Index} \quad = \frac{1}{N}\sum_{i=1}^{N} |d_i|$$

Where N is the total number of individuals, and $d_i = \text{Rank in Initial Period} - \text{Rank in Subsequent Period}$

$$\text{Upward Mobility U} = \frac{1}{N}\sum_{i=1}^{N} max(d_i, 0)$$

$$\text{Downward Mobility D} = \frac{1}{N}\sum_{i=1}^{N} max(-d_i, 0)$$

We map the Lerman-Yitzhaki Upward Mobility and Downward Mobility across each district, probing whether the spatial clusters we previously identified disclose any geographic associations with wealth transitions. This examination unveils distinct and significant networks. Within each spatial cluster, we

construct unique maps that reveal new geographical networks, shaped by shared and analogous human development realities. These clusters, akin to states or administrative level 1 boundaries, encompass districts where we observe particular epicenter(s) or nodes of social mobility.

3. Exploratory and Empirical Results

Educational Participation Indicators: In the context of numerous educational indices such as PISA, GER, and MDG, we developed a comprehensive metric to gauge educational participation across individual, household, and district strata levels. The raw measurement, referred to as E_{raw} , represents the level of educational participation, with higher values signifying lower participation. To create a more intuitive understanding, we normalize this metric by dividing the raw score by 4, which provides us the Individual Level Educational Participation Deficit Indicator (E) and the inverse (1 - E) is termed the Individual Educational Participation Score.

For children aged 5 to 12 years:

- $\boldsymbol{E_{raw}}$ **= 4** if the child is not currently attending school.
- $\boldsymbol{E_{raw}}$ **= 0** if the child is currently attending school or has attended school at some point during the survey year.

For adolescents aged 12 to 18 years:

- $\boldsymbol{E_{raw}}$ **= 4** if the adolescent is not attending school and has less than six years of schooling.
- $\boldsymbol{E_{raw}}$ **= 3** if the adolescent is not attending school but has six or more years of schooling, though has not completed secondary education.
- $\boldsymbol{E_{raw}}$ **= 2** if the adolescent is attending school, has six or more years of schooling, but still pursuing primary education
- $\boldsymbol{E_{raw}}$ **= 1** if the adolescent is not attending school but has completed secondary education.
- $\boldsymbol{E_{raw}}$ **= 0** if the adolescent is attending school, has started or completed secondary education but not higher education, and has six or more years of schooling.

For individuals aged 18 years and above:

- $\boldsymbol{E_{raw}}$ **= 4** if the individual has less than six years of schooling.
- $\boldsymbol{E_{raw}}$ **= 3** if the individual has six or more years of schooling, completed primary education but has not pursued secondary education, and is currently not out of school.
- $\boldsymbol{E_{raw}}$ **= 2** if the individual has six or more years of schooling, has started secondary education but has not completed it, and is currently not attending school.
- $\boldsymbol{E_{raw}}$ **= 1** if the individual has six or more years of schooling, has completed secondary education but not higher education, and is currently not attending school.
- $\boldsymbol{E_{raw}}$ **= 0** if the individual is either attending school and has completed secondary education or higher, or is not attending school but has pursued higher education.

The Household Level Educational Participation Deficit Indicator is calculated by averaging the normalized E values for all individuals within a household. The inverse of this household-level value provides a household participation score. At the district level, the Educational Participation Deficit Index is determined by averaging the Household Level Educational Participation Deficit Indicators for all households or individuals within a district

For example, for a 10-year-old child who is currently attending school, the E_{raw} value is 0. Normalizing this, we get E = 0, and the Individual Educational Participation Score is 1 - 0 = 1(best). A 14-year-old adolescent who is not attending school and has less than six years of schooling would have an E_{raw} value of 4 . When normalized, E = 4 / 4 = 1, making the Individual Educational Participation Score 1 - 1 = 0(worst). A 15-year-old adolescent who is not attending school but has six or more years of schooling, though has not completed secondary education, has an E_raw value of 3. Normalizing this, we get E = 3 / 4 = 0.75, and the Individual Educational Participation Score is 1 - 0.75 = 0.25. A 20-year-old individual who has six or more years of schooling, has completed secondary education but not higher education, and is currently not attending school, has an E_{raw} value of 1. Normalizing this, we get E = 1 / 4 = 0.25, and the Individual Educational Participation Score is 1 - 0.25 = 0.75.

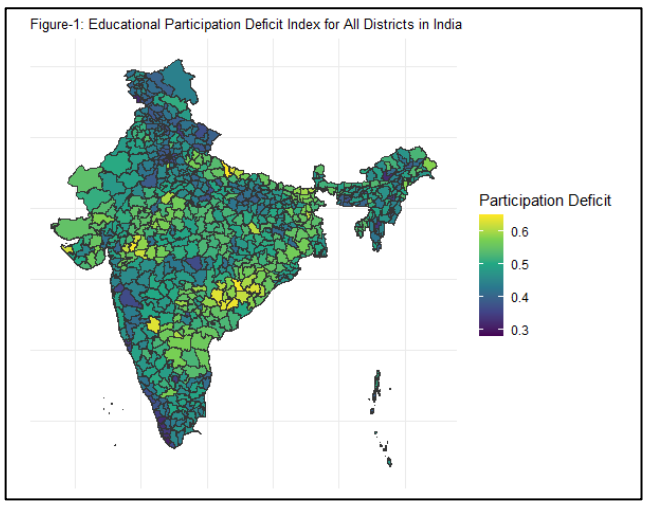


Figure-1: Educational Participation Deficit Index for All Districts in India

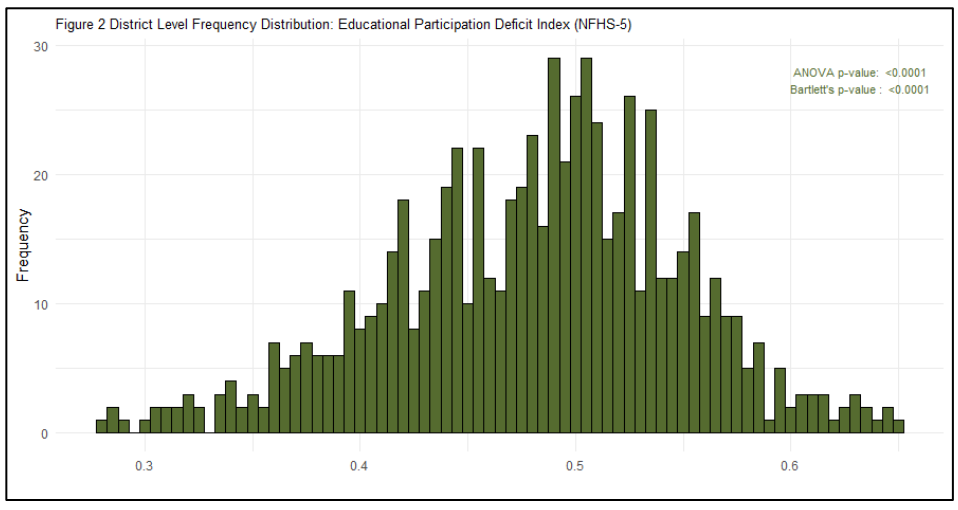
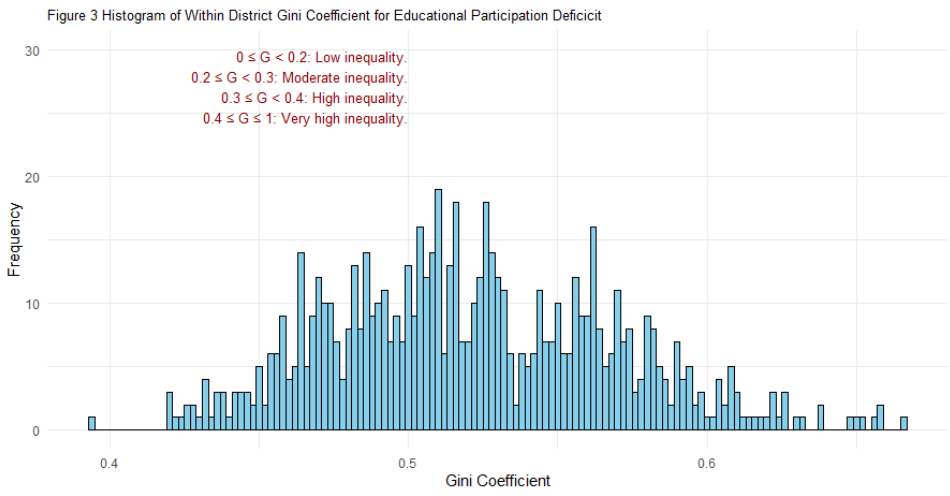


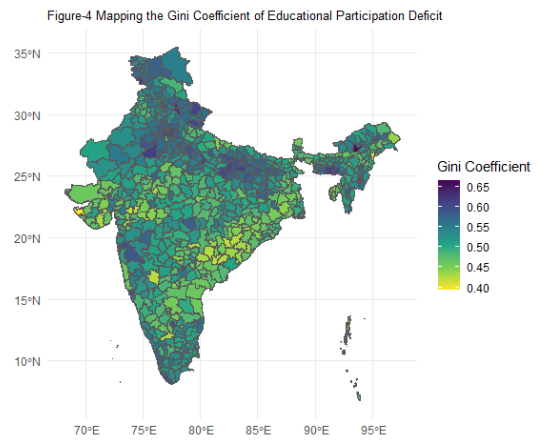
Figure 2 District Level Frequency Distribution: Educational Participation Deficit Index (NFHS-5)

In Figure 1 , we observe that the distribution of educational participation exhibits spatial clustering, where neighboring districts manifest similar participation patterns. This phenomenon is starkly illustrated by the striking disparities between major urban centers and more marginalized districts. Mumbai and Chennai, for instance, exhibit participation deficit indices of 0.2783 and 0.2845, respectively, situating them far ahead of Bahraich in Uttar Pradesh, which registers the most severe educational participation deficit in the country at 0.6509. Nabarangapur in Odisha follows closely with a deficit of 0.6457693. Such disparities compel us to examine the spatial clusters and to interrogate whether the primary mechanisms of educational participation deficits also exhibit similar clustering. Prior to our spatial analysis, we will present additional summary statistics, which underscore that heterogeneity and variability are not confined to inter-district comparisons but permeate within districts and across communities.

Data : In the study we have worked with IPUMS-DHS 2019-21 [India] data which is the National Family Health Survey-5 [NFHS-5] data and it is the largest sample survey conducted in India to date. The National Family Health Survey 2019-21 (NFHS-5), the fifth in its series, provides comprehensive data on population, education, and nutrition for India, its states, union territories (UTs), and 707 districts. Conducted under the stewardship of the Ministry of Health and Family Welfare (MoHFW) and coordinated by the International Institute for Population Sciences (IIPS), NFHS-5 was funded by MoHFW and received technical assistance from ICF, USA. The survey, executed in two phases, amassed information from 636,699 households, 724,115 women, and 101,839 men. NFHS-5 builds on the foundations of previous surveys, starting from NFHS-1 in 1992-93, maintaining continuity in content and methodology while progressively expanding its scope. NFHS-4 (2015-16), with its enhanced sample size, provided district-level data and ensured comparability with earlier rounds. Similar to NFHS-4, NFHS-5 delivers district-level estimates for numerous crucial indicators, including the wealth index, which is extensively utilized in this study. It also introduces the new topic of preschool education, while continuing to cover incomplete education, access to schooling, reasons for drop-out, and occupation history. These elements enable a thorough analysis of educational participation and its socio-economic implications at both district and national levels, as well as the spatial spillover of educational barriers. The NFHS-5 sample design aimed to produce reliable data at the district, state/UT, and national levels. A uniform, stratified two-stage sample design was adopted, with districts stratified into urban and rural areas. Rural strata were further sub-stratified based on village population and the percentage of the population belonging to scheduled castes and scheduled tribes (SC/ST). Villages and Census Enumeration Blocks (CEBs) were chosen as Primary Sampling Units (PSUs), sorted by women's literacy rates in rural areas and by SC/ST population percentages in urban areas before selection. Rural villages were selected with probability proportional to size (PPS), and each rural stratum was divided into six approximately equal substrata. Urban CEBs were sorted similarly, and PSUs were selected using PPS systematic sampling. Each PSU or segment of a PSU had an estimated 100-150 households, with 22 households per cluster selected through systematic sampling. A total of 30,456 PSUs were selected across the country, with fieldwork completed in 30,198 PSUs. Four survey questionnaires—Household, Woman, Man, and Biomarker—were translated into 18 local languages and administered using Computer-Assisted Personal Interviewing (CAPI). Data was collected on household demographics, socio-economic characteristics, health insurance coverage, digital banking, Internet usage, land ownership, and mosquito net usage, among other topics. The survey also included extensive training and quality control measures. Training of Trainers (ToT) workshops were conducted for field coordinators, who then trained fieldworkers at the state/UT level. Fieldwork was monitored by multiple levels of supervisors, including district coordinators, IIPS project officers, and senior staff from Field Agencies. Data quality was ensured through daily data transfers to IIPS, extensive data quality checks, and real-time feedback to field teams. NFHS-5 achieved high response rates, with 98% of selected households successfully interviewed, 97% response rate for women, and 92% for men. The data collected provide valuable insights into India's health and family welfare landscape, assisting policymakers and program managers in setting benchmarks and assessing progress. This extensive and detailed survey is the largest sample survey conducted in India to date, offering critical data to inform public interventions and policy decisions.

3.1 Summary Statistics :





Contemporary literature on education in Indian districts underscores the profound interplay of socio-economic factors, including caste dynamics, on educational outcomes. Studies such as those by Desai et al. (2010) and Dreze and Sen (2013) have documented the entrenched inequalities in educational attainment linked to caste and socio-economic status. These disparities are not merely reflections of economic inequities but are also perpetuated by social and cultural barriers. For instance, children from Scheduled Castes (SC) and Scheduled Tribes (ST) navigate a labyrinth of challenges in accessing quality education, ranging from discrimination within schools to a lack of educational resources in their communities . Barriers to education in these contexts are multifaceted. The impact of these barriers is particularly pronounced in rural areas where economic constraints, gender discrimination, inadequate infrastructure, and insufficient teacher training are critical factors that exacerbate educational inequalities. Research by Banerjee and Duflo (2011) highlights how economic hardships limit educational opportunities for children in impoverished households, leading to a cycle of disadvantage. In Table-1, we have provided a summary across different demographic age groups, rural/urban positions, gender, and caste, assessing the statistical significance where we found most tests to show significance.

The histogram (Figure -3) of the Gini coefficient for educational attainment across districts showcases a broad spectrum of inequality levels and underscores that almost all districts display significant internal heterogeneity, irrespective of overall participation rates. Figures 3 and 4 highlight the overall heterogeneity in educational attainment within districts. The Gini coefficient, a measure of statistical dispersion, represents income or wealth inequality within a nation or social group. It ranges from 0 to 1, where 0 indicates perfect equality (everyone has the same income or wealth) and 1 indicates perfect inequality (one person has all the income or wealth, and everyone else has none). This indicator can also be used to measure inequality in other distributions, such as educational attainment or health outcomes . A Gini coefficient of 0.4 or above is generally considered as very high inequality.

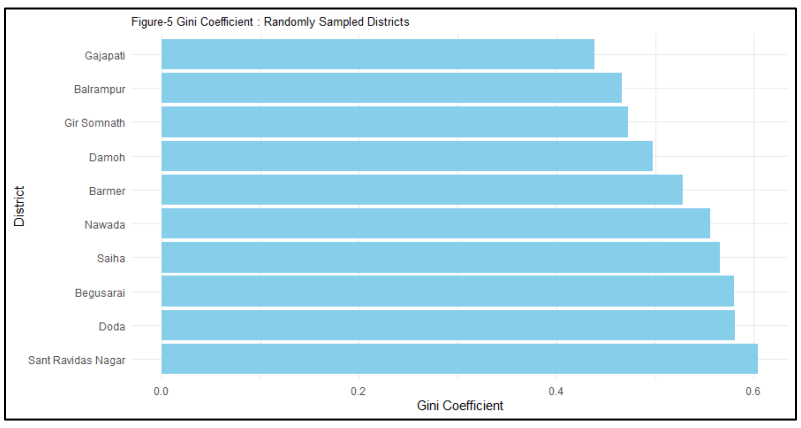


Figure-5 further elucidates this point by showcasing a random sample of districts with random levels of within-district educational inequality, emphasizing that most districts might have need for targeted interventions.

Jalan and Ravallion (2002) noted that geographic disparities in education are closely linked to broader patterns of regional development and poverty. But, in our data we see notable concentration of districts experiencing moderate to high inequality (Gini coefficient between 0.3 and 0.5) irrespective of overall participation rates (as observed in Figure 1). So, it will be reasonable to explore whether regions with higher participation deficits always correspond to areas with longstanding socio-economic disadvantages and even if it does, there will certainly be other principal actors. For instance, Mumbai and Chennai, which have high educational participation and low deficit indices of 0.2783 and 0.2845, respectively, exhibit Gini coefficients of 0.638 and 0.647, signalling substantial inequality within these highly participative districts. Conversely, Nabarangapur in Odisha and Bahraich in Uttar Pradesh, which suffer from the worst participation deficits at 0.6457693 and 0.6509, respectively, display Gini coefficients of 0.4207 and 0.4467. This is likely because there is limited room for variability at the lower end of the threshold. However, this suggests that even developed urban centres with high participation can exhibit significant within-district inequality . Spatial clustering of educational participation deficits, as indicated by the map in Figure 1, also finds some reflection in Figure 3. While high participation clusters may not coincide with high inequality clusters, the presence of these clusters points to regional patterns of inequality likely stemming either from historical, economic, and policy-driven factors, as noted by Jalan and Ravallion (2002), or from the fact that demographic communities facing the highest discrimination often come from regions where there may be national-level oversight or local discrimination affecting overall attainment. Research by Kingdon (2007) and Tilak (2007) supports the notion that educational inequalities in India are deeply rooted in socio-economic and cultural factors. Kingdon's study highlights the role of gender and caste in shaping educational outcomes, while Tilak emphasizes the need for substantial public investment in education to bridge these gaps. These insights underscore the multifaceted nature of educational disparities and the necessity for targeted interventions that address both economic and social dimensions of inequality. Along with gender and caste, in this study we also explore the disparity between different religious groups.

In conclusion, within-district educational participation deficits underscore the significant and varied challenges to achieving educational attainment in India, as a part of Minimum Development Goals. Addressing these disparities requires an unbiased approach that considers the socio-economic and cultural contexts of different regions and tests existing hypotheses regarding both the demand and supply sides of education for marginalized communities. This aligns with the broader literature advocating for comprehensive strategies to bridge educational gaps and promote inclusive growth. Figures 3 to 6 provide a visual representation of the complex landscape of educational inequality in India, emphasizing both within and between district disparities. In Table-1, we will see the statistical significance and exact figures regarding these disparities as part of our initial exploratory analysis in elaborate manner.

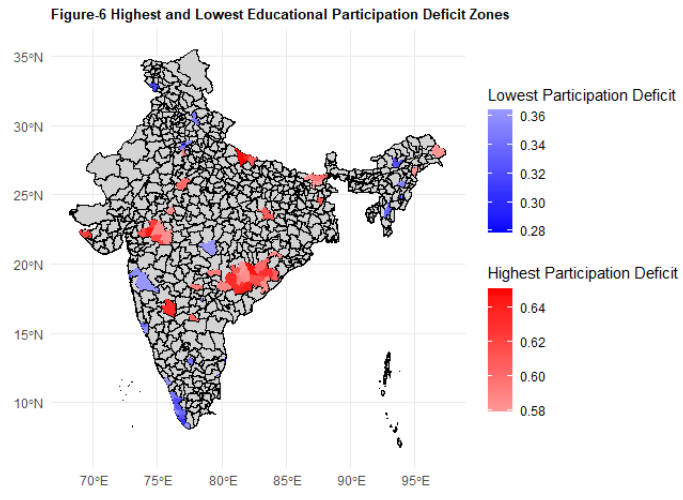


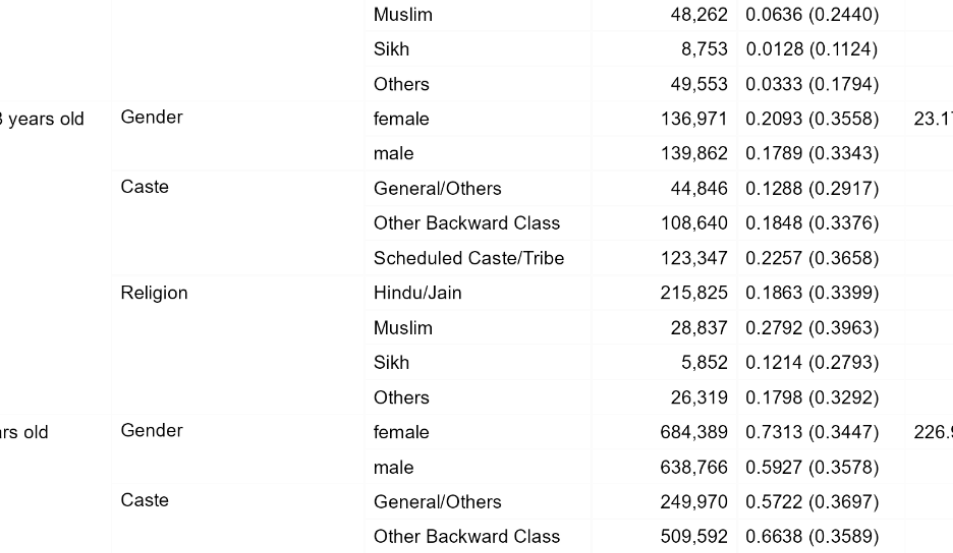
Table 1: Average Individual Educational Participation Deficit Scores Across Demographic Group

Table 1: Average Individual Educational Participation Deficit Scores Across Demographic Groups							
Urban/Rural	**Age Group**	**Demographic Parameter**	**Demography**	**Sample Size**	**Mean (SD)**	**t value**	**F value**
Urban	Below 12 years	Gender	female	56,857	0.0256 (0.1580)	0.77	
			male	62,406	0.0249 (0.1559)		
		Caste	General/Others	31,031	0.0190 (0.1365)		34.896***
			Other Backward Class	50,416	0.0267 (0.1611)		
			Scheduled Caste/Tribe	37,816	0.0285 (0.1665)		
		Religion	Hindu/Jain	83,874	0.0220 (0.1468)		127.921***
			Muslim	23,017	0.0428 (0.2025)		
			Sikh	2,179	0.0133 (0.1146)		
			Others	10,193	0.0147 (0.1204)		
	12 to 18 years old	Gender	female	37,903	0.1222 (0.2882)	-6.88***	
			male	41,300	0.1367 (0.3035)		
		Caste	General/Others	21,090	0.0915 (0.2536)		267.047***
			Other Backward Class	33,127	0.1360 (0.3024)		
			Scheduled Caste/Tribe	24,986	0.1538 (0.3179)		
		Religion	Hindu/Jain	55,995	0.1110 (0.2765)		610.109***
			Muslim	14,817	0.2218 (0.3673)		
			Sikh	1,581	0.1013 (0.2653)		
			Others	6,810	0.0902 (0.2453)		
	18+ years old	Gender	female	227,164	0.5302 (0.3892)	99.359***	
			male	221,589	0.4191 (0.3588)		
		Caste	General/Others	134,729	0.3816 (0.3663)		6386.097***
			Other Backward Class	184,729	0.5025 (0.3772)		
			Scheduled Caste/Tribe	129,295	0.5350 (0.3751)		
		Religion	Hindu/Jain	335,547	0.4591 (0.3791)		2466.669***
			Muslim	66,356	0.5893 (0.3686)		
			Sikh	10,557	0.4492 (0.3751)		
			Others	36,293	0.4266 (0.3547)		
Rural	Below 12 years	Gender	female	219,433	0.0379 (0.1909)	3.395***	
			male	235,255	0.0360 (0.1864)		
		Caste	General/Others	70,425	0.0244 (0.1542)		332.345***
			Other Backward Class	174,149	0.0335 (0.1799)		
			Scheduled Caste/Tribe	210,114	0.0440 (0.2051)		
		Religion	Hindu/Jain	348,120	0.0344 (0.1821)		397.699***
			Muslim	48,262	0.0636 (0.2440)		
			Sikh	8,753	0.0128 (0.1124)		
			Others	49,553	0.0333 (0.1794)		
	12 to 18 years old	Gender	female	136,971	0.2093 (0.3558)	23.173***	
			male	139,862	0.1789 (0.3343)		
		Caste	General/Others	44,846	0.1288 (0.2917)		1370.437***
			Other Backward Class	108,640	0.1848 (0.3376)		
			Scheduled Caste/Tribe	123,347	0.2257 (0.3658)		
		Religion	Hindu/Jain	215,825	0.1863 (0.3399)		727.232***
			Muslim	28,837	0.2792 (0.3963)		
			Sikh	5,852	0.1214 (0.2793)		
			Others	26,319	0.1798 (0.3292)		
	18+ years old	Gender	female	684,389	0.7313 (0.3447)	226.917***	
			male	638,766	0.5927 (0.3578)		
		Caste	General/Others	249,970	0.5722 (0.3697)		12371.23***
			Other Backward Class	509,592	0.6638 (0.3589)		
			Scheduled Caste/Tribe	563,593	0.7062 (0.3435)		
		Religion	Hindu/Jain	1,044,353	0.6646 (0.3593)		478.747***
			Muslim	109,712	0.6942 (0.3542)		
			Sikh	35,643	0.6240 (0.3554)		
			Others	133,447	0.6499 (0.3478)		

Note: The table shows the demographic 'Differences in Partcipation Deficit' by gender, religion, and caste.

One notable divergence from the existing findings in the NAS (National Achievement Survey 2023) report, which demonstrated that girls outperformed boys in learning outcomes and enrollment levels, became apparent when we considered enrollment data from the much larger NSSO sample. When we distinguish between urban and rural environments, the numbers we uncovered showed that, in most cases, the participation rate of girls was lower than that of boys contrary to the NAS report which perhaps overrepresented the urban households. In Table 1, we see an inverse nature of the participation deficit between boys and girls in the age group 12-18 years old: In urban areas, likely to host private schools, girls exhibit higher enrollment rates than boys. In contrast, in rural regions, predominantly served by public schools, boys' enrollment surpasses that of girls. This discrepancy between enrollment, when viewed through the lens of age-specific grouping and graded coding of educational participation, illuminates the disuniformity of student attrition between successive grades.

Figure 7



In Fig -7 we can discern the overall frequency counts of statistically significant and non-significant differences when we look at each district separately and we observe the skewedness of this distribution . In our exploratory analysis above, we have charted the landscape of educational participation across districts by presenting district-wise overall participation rates. This offers a granular view of educational engagement. Additionally, we have calculated and depicted the Gini Coefficient for each district, a metric that elucidates

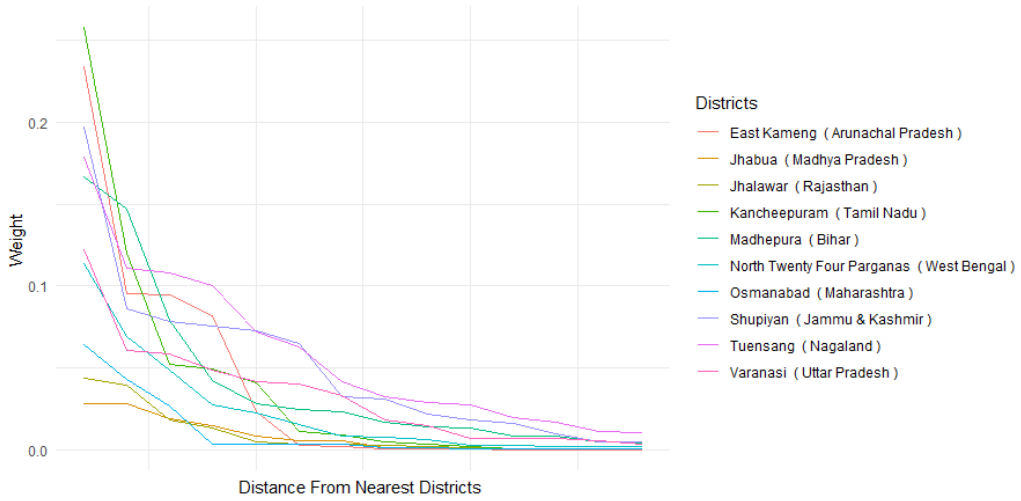
the inequality in educational participation at the district level. By plotting the distribution of these Gini Coefficients, we have highlighted regions with significant disparities, thereby enhancing our understanding of the intensity of local barriers to participation (Fig 2). To provide context to our findings, we have incorporated specific examples from over 700 districts, casting light on familiar places so that readers can relate to known districts and understand the state of educational participation in their regions (Fig 1, Fig 4). The focus of our analysis has been the discrepancies between demographic groups—male versus female, SC/ST/OBC versus General/upper caste, and Muslims versus other religious groups. Our findings reveal stark contrasts in educational participation. The urban-rural divide is also critical to examine in this context of inequality (Table 1). While district-level statistics offer granular insights, understanding the overall inclination at the national level necessitates examining the ratio of high and low-performing districts concerning the respective demographic groups. To this end, we have conducted district-wise t-tests and presented the results using dot plots (Fig-7).

In our subsequent analyses, we will sustain the district-level focus of our investigation; however, the novel question we will address pertains to the influences exerted by neighboring districts, an inquiry that mandates the utilization of spatial statistical methods. In clusters where neighboring districts consistently underperform, the necessity for targeted interventions becomes glaringly obvious. Exploratory work or general regression methods are insufficient to capture these intricate spatial effects. We will first illustrate that such spatial effects exist and are significant. Therefore, before delving into analyzing the returns on education and investigating potential evidence of active discrimination once education is controlled for, to achieve true unbiasedness, it will be crucial to obtain estimates that also account for the influence of neighboring districts across different demographic groups. In the following sections, we will employ a range of spatial statistical methods to present the discussion with theoretical sections as needed, ensuring that theory, examples, and results are presented side by side to make the information easily accessible and comprehensible for the reader.

3.2 Spatial Exploratory Data Analysis :

Moran's I statistic is a fundamental tool in spatial analysis, used to measure spatial autocorrelation —that is, how much a variable is correlated with itself across a given space. In our framework it is used to identify whether we have evidence for spatial autocorrelation between participation rates at neighboring districts i.e. whether spatially proximate social units have spillover effects on mutual supply and demand of education.

Figure 8 Sample Weight Assignments for 10 Randomly Selected Districts

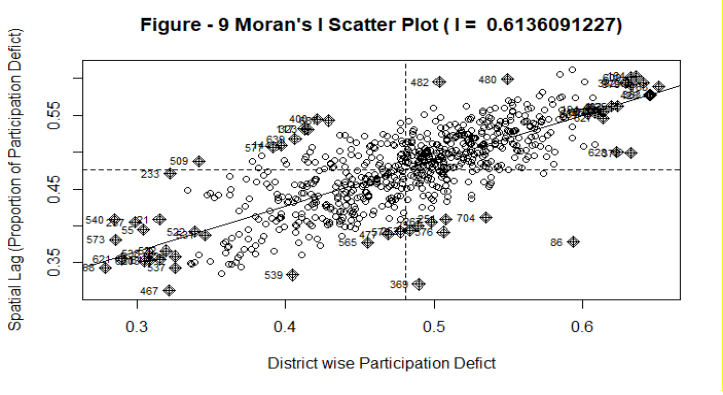


Moran's I for **Geographic Disparity:**

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

w_{ij} *:* Spatial weight between ith and jth district , $\bar{x}$ *:* Mean across all locations

Selecting an appropriate weighting matrix has various potential options where nearest neighbors are always assigned higher weights. Generally, different types of weighting matrices can be considered. In our study, the initial assumption posits that spatial effects are not confined by state (admin 1 level) boundaries, thereby a straightforward 707by707 weighting matrix can be sufficient, at least initially. For our study, *we define the* weight formula as, $w_{ij} = \exp\left(-\frac{\text{distances}}{\text{threshold}}\right)$ with a threshold of 0.2. This chosen threshold brings the weight to zero approximately after the nearest 7th district (depending on how large and distant the districts in that cluster are) ensuring that distant influences are minimized exponentially . Figure 8 illustrates this rate of decay i.e. what a threshold of 0.2 means, and how the influence diminishes with increasing distance. For example, weights assigned to Jhalwar (Rajasthan), where neighboring districts are larger, diminish sharply and decay to zero approximately after the 5th district. In contrast, for Tuensang (Nagaland), the weight does not decay to zero even after the 15th district because the northeastern part of India has smaller districts. To note, this weight structure is a general function of distance. When applied to our dataset, it reveals the spatial lags and global autocorrelation between participation rates (Figure 9). The Moran's I value is positive and significantly high (Moran's I = 0.6136, p-value < 2.2e-16). This is our statistical evidence that educational participation is not randomly distributed but follows a pattern influenced by spatial proximity.



In Figure 9, on the X-axis, we plot the rates of participation deficit for each district, while the Y-axis represents the spatial lag. This visualization shows the spatial clustering of educational participation rates at a global level.

The local spatial correlation is given by:

$$LISA_I = \frac{x_i - \bar{x}}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \sum_{j \neq i} w_{ij}(x_j - \bar{x})$$

LISA stands for Local Indicators of Spatial Association.

The observations can be classified into 4 categories based on the LISA I values. A positive LISA I signifies that a district is surrounded by others with similar educational characteristics, creating clusters of homogeneity (Fig. 10). Conversely, a negative LISA I indicates points of discontinuity, where a district is encircled by districts exhibiting different states of education, thus identifying these as spatial outliers [Fig 12]. These outliers are further distinguished as HOT or COLD spots, contingent upon whether their educational participation rates are significantly higher or lower than the average.

Figure 10 LISA Map (Positive-Positive & Negative-Negative)

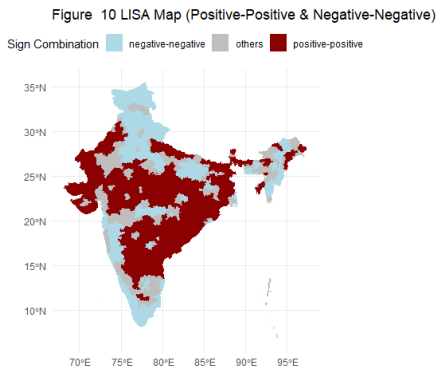


Figure 11 LISA Cluster Map

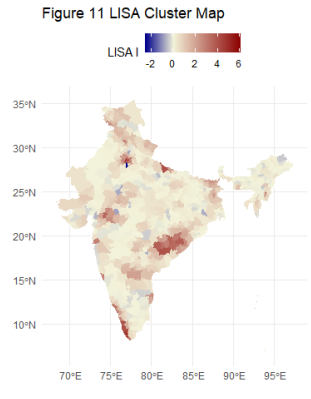
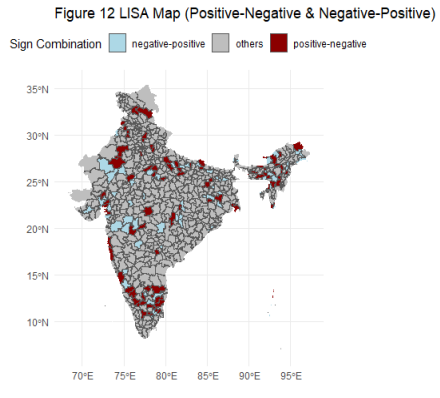


Figure 12 LISA Map (Positive-Negative & Negative-Positive)



3.3 Empirical Analysis : District Level Mixed-Effect Regressions

In this section, we decompose the total variabilities of educational participation at each district level :

$$\textit{Number of Years Education}_{ijd} = \beta_{0d} + \sum_k \beta_{kd} \cdot (\textit{demographic\_factor}_k) + \sum_p \beta_{pd} \cdot \left(\textit{Gender} \times \textit{demographic\_factor}_p\right) + \underbrace{\beta_{6d} \cdot exp\left(\frac{-1}{age_{ijd}^2}\right)}_{\text{represents the secondary schooling dropout}} + u_{jd} + \epsilon_{ijd}$$

This is our regression model for the d^{th} district. We perform 707 district-level regressions for 707 districts of India: $d = 1, 2, \dots, 707$

Here,

$u_{jd} \sim N(0, \sigma_u^2)$, Household level random effect for j$^{\text{th}}$ Household at the d^{th} district.

$\text{Number of Years Education}_{ijd}$, Number of years of education for i^{th} individual from j^{th} Household at the d^{th} district.

(We focused on an age group ranging from 5 to 25 years old)

$\epsilon_{ijd} \sim N(0, \sigma^2)$

$k \in \{Muslim, SC/ST, Female\}$, $p \in \{Muslim, SC/ST\}$

Some districts did not meet the full-rank requirements, and when the number of observations from a particular demographic subset is very low, the estimates lack statistical power and meaningfulness. For instance, in Kupwara, Jammu & Kashmir, we did not have enough female SC/ST observations. In such cases, we excluded that demographic category as a cofactor in the district-specific regression and also excluded the respective interaction effect.

Our models had a median R^2 of 0.7412, with the 1st and 3rd quartiles at 0.6956 and 0.7932, respectively.

Figure 13 Conditional(random+fixed) R-squared for All Distrcits

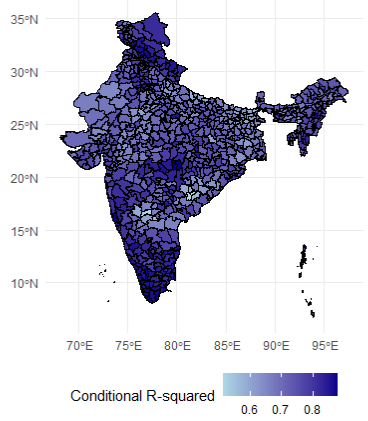
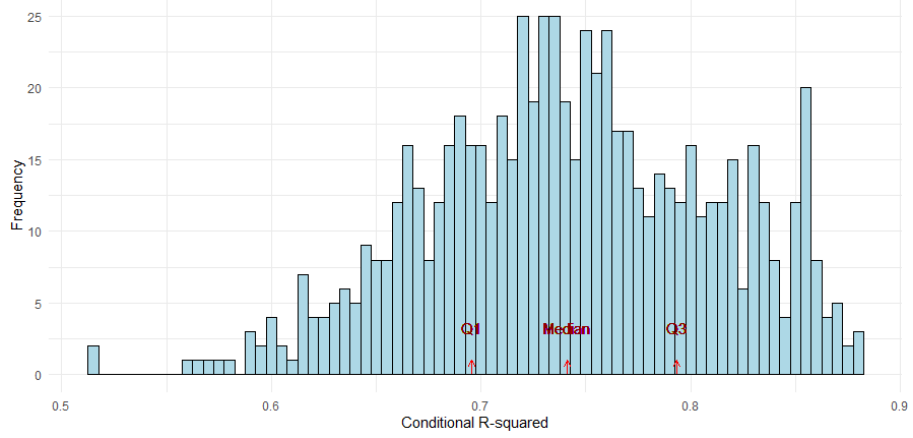


Figure 14 Histogram of Conditional(random+fixed) R-squared



Our model describes a significant proportion of total variability

The β Estimates: While this model captures a significant percentage of variability, our aim is not predictive analysis. Rather, we seek to interrogate the distribution of the coefficients for Female (i.e., when the individual is female; female = 1), Muslim (i.e., when the individual is muslim; muslim = 1), and SC/ST (i.e., when the individual is SC/ST; SC/ST = 1). We focus on discerning the districts where these estimates hold significance and identifying the hot points within these spatial distributions. In Fig-15-17 we show the main-effects (i.e. not the interaction effects):

Figure 15 Districts with Significantly Negative Female Educational Participation

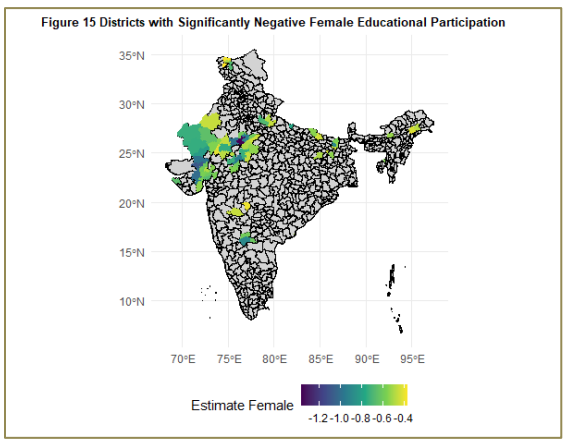


Figure 16 Districts with Significantly Negative Educational Participation of Muslim Community

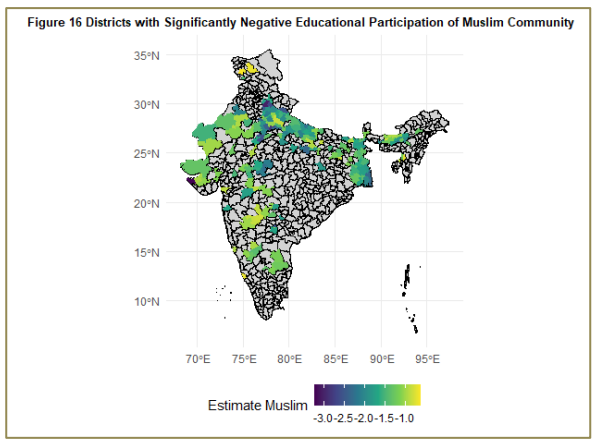
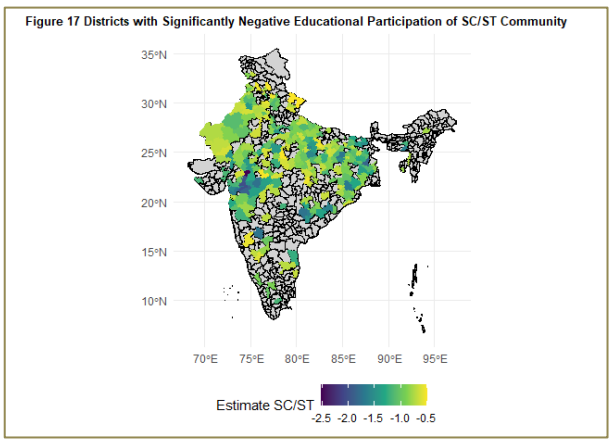


Figure 17 Districts with Significantly Negative Educational Participation of SC/ST Community



Among the districts with adequate observations, we identified:

a) Female participation in education for the age group of 5 to 25 was notably low in 9.47% of districts (67 out of 707 districts), with a p-value of less than 0.01.

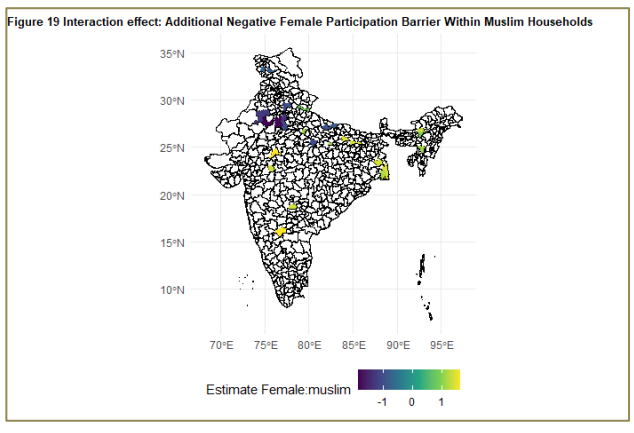
b) For the same age group in Muslim communities, 53.31% of districts (185 out of 347 districts) showed similarly low participation rates. (in rest of the districts we did not have adequate Muslim responders)

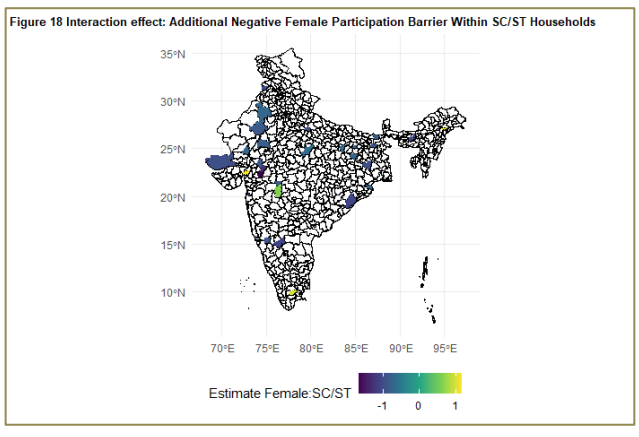
c)For SC/ST groups, this low participation was observed in 42.56% of districts (275 out of 646 districts).

- **Note :** Figures 15 to 17 display the regression summary results (magnitude and p-values) of the 707 regressions we conducted. In the maps, the main-effect coefficients **($\beta_{female,d}$, $\beta_{SC/ST,d}$ and $\beta_{muslim,d}$) for districts *d= 1, 2… 707*** are provided for each district where the coefficients were statistically significant. If a particular β estimate is greyed out on the maps, it indicates that no significant result was observed for that demographic cofactor in the respective district. Mapping the coefficients is the most efficient way to summarize the results, with the overall regression efficiency shown in the histogram in Figure 14 (as well as Figure 13).

- **Do Muslim or Dalit (SC/ST) women encounter greater barriers to participation due to regressive attitudes within households in backward communities?**

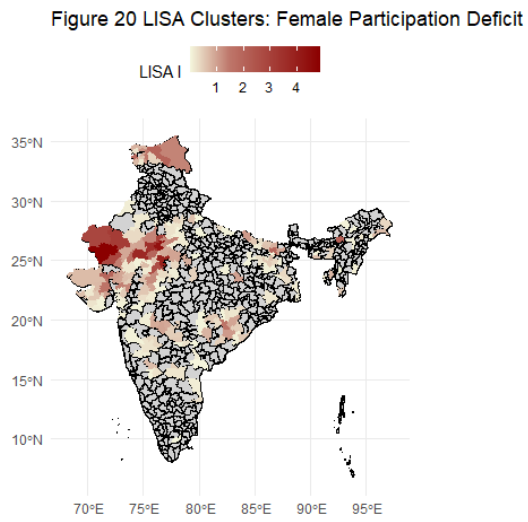
  We examine the interaction effects, which represent the non-linear component beyond the additive effects of the two factors Female and Muslim/SC/ST:

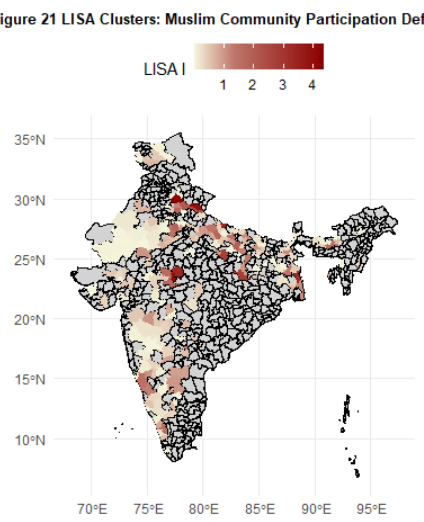


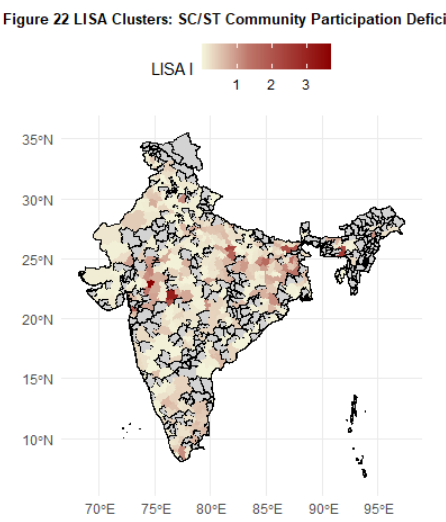


  This indicates that the common belief that Muslim and SC/ST households are particularly regressive towards women is not justified, at least with respect to the barriers to educational participation in India, based on the data that we have collected in 2019-21.

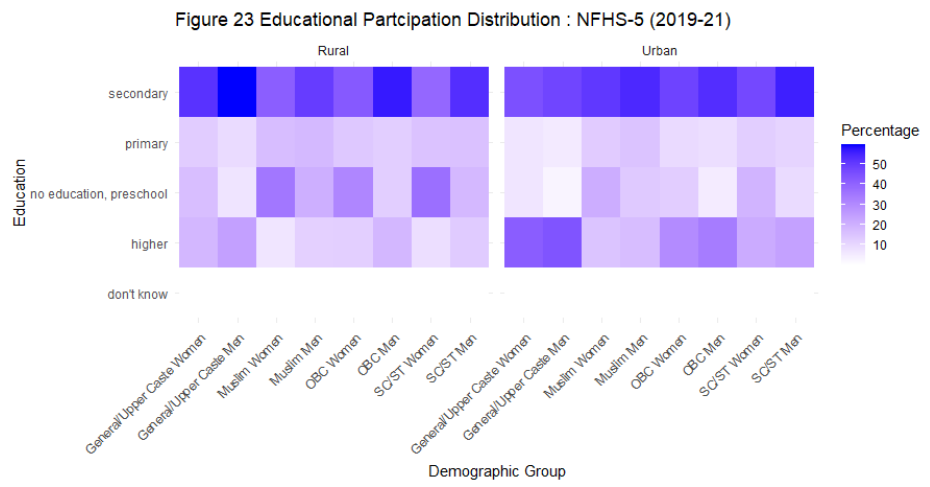
- **Spatial Spillover (The Local Clusters/hot spots using LISA) :** Finally, we examined Local Indicators of Spatial Association (LISA) to identify the districts that exert strong influence

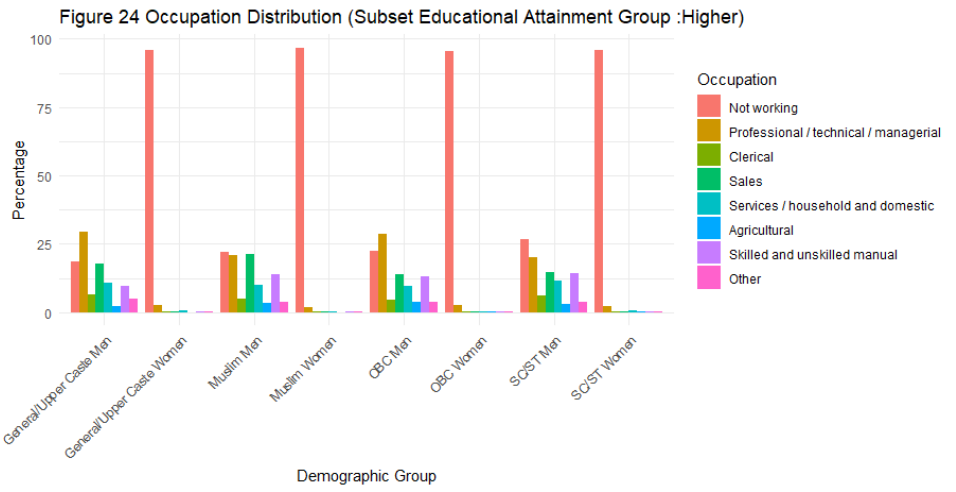




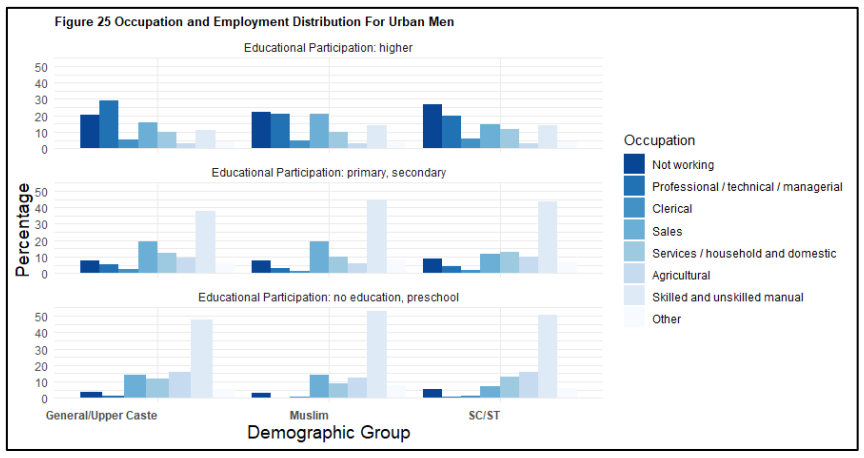


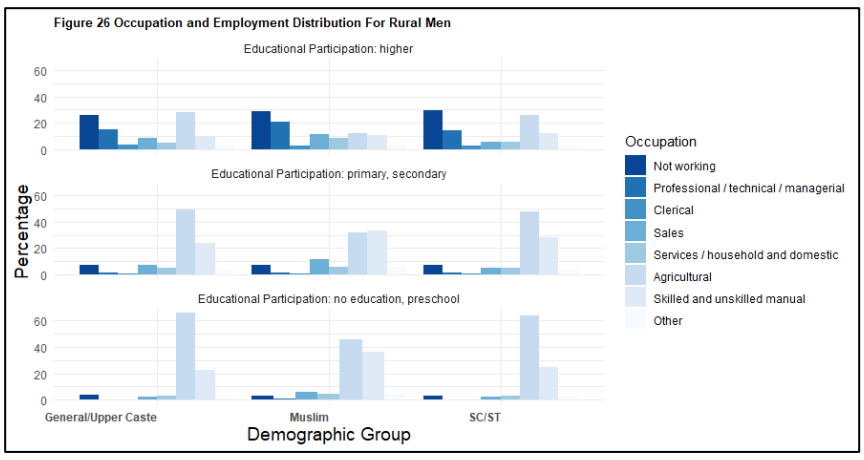
3.4 Does educational participation provide equal benefits across different genders and communities?

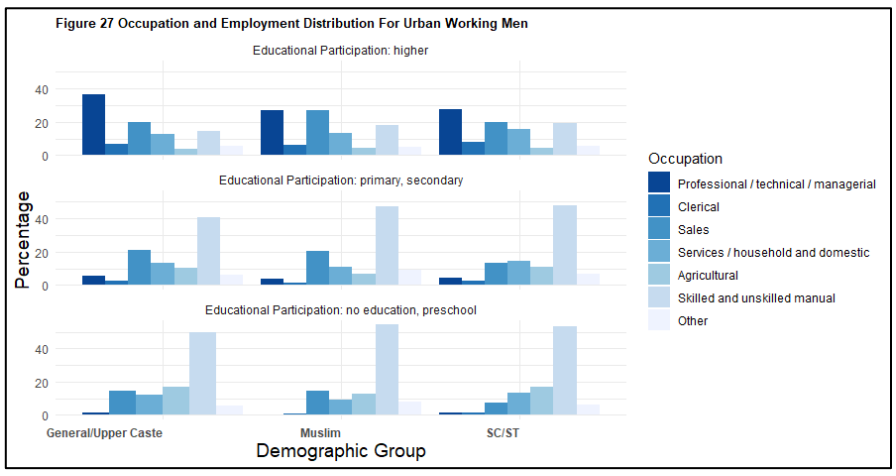




According to the World Bank, a one-year increase in the average years of schooling can elevate a country's GDP growth by 0.37%. Similarly, a 1% increase in the literacy rate can enhance GDP growth by 0.3%. "Pre-market endowment" encompasses the array of skills, education, and other attributes—such as overall health, physical condition, social and cultural capital, and personal characteristics like intelligence and motivation—that individuals possess before entering the labour market. These pre-market endowments critically shape job opportunities, earning potential, and career trajectories. In this section, we seek to interrogate the extent to which education alone can act as a sufficient or predominant catalyst for economic growth within this diverse nation. Furthermore, we examine the junctures and demographic contexts where additional, targeted interventions may be imperative to address the complexities of systemic inequities.







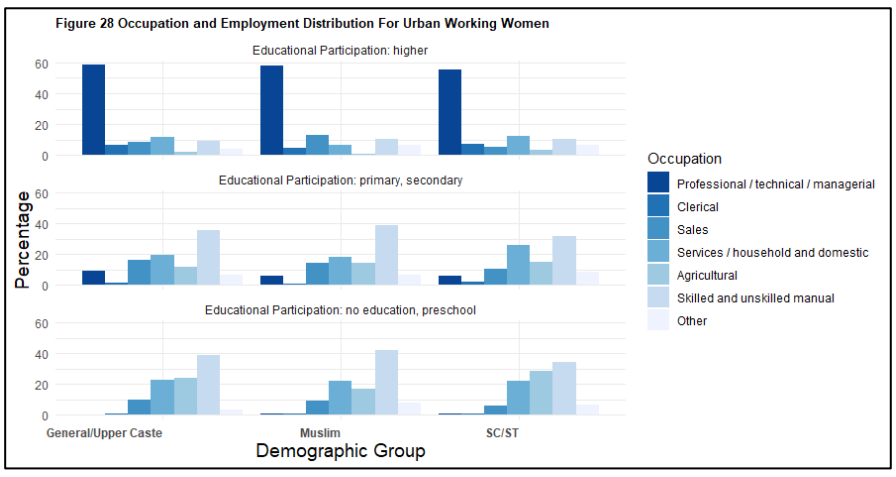


Fig -25 and Fig - 26 show a comparison between urban (Fig 25) and rural (Fig 26) men who are 18 years old or older. The comparison looks at different education levels and how education impacts job clustering. As men from different demographic groups progress from having no education to primary/secondary education and then to higher education, the graphs illustrate how they tend to group into different occupations and the differences in returns to pre-market educational attainment for the three demographic groups. In contrast, Figures 27 and 28 focus on urban working women and men. It is certainly troubling that many women are not part of the workforce despite higher education (as shown in Figure 23 & Figure 24), and it is important to note that a considerable number of men, particularly in rural India, with higher education are also out of the workforce. But in the current analysis we also aim to discern the persistent gender discrepancy in the labor market i.e., the extent to which educational attainment benefits men over women, even among those who are employed.

Our observations align with ordinary understanding and corroborate the district-level analysis previously conducted in Section 2:

- With education, the likelihood of securing professional, technical, or managerial positions increases for all groups.
- Social and cultural capital plays a significant role, as men from the general category tend to secure more sophisticated and higher-paying jobs compared to Muslim and SC/ST men. This suggests the presence of discrimination or barriers for Muslims and SC/STs, as their occupational clustering differs despite having similar educational attainment, skills, and training. In Table 2 (at the end of this initial discussion), we have provided the exact values.

- Learning outcomes may vary between rural and urban areas. This could be due to differences in learning outcomes or a lack of suitable job opportunities and awareness of job information and trends.
- Among urban men and women with higher education, men predominantly work in professional, managerial, and technical jobs, while women are more likely to work in sales as well as professional roles.
- For all groups, the percentage of unskilled manual laborers decreases as they progress from having no education or only secondary education to obtaining higher education.
- Additionally, one specific and intriguing observation here might contradict the conventional expectation and perhaps may also be a barrier to promoting educational participation. In Figures 3 and 4, we observe that while higher education should theoretically facilitate a shift from unskilled manual or agricultural labor to advanced technical roles for both rural and urban men, it concurrently elevates the risk of unemployment. We see, rural men who transitioned from agricultural work to higher education, most likely face joblessness if they fail to secure advanced positions. Similarly, urban men previously employed in unskilled manual jobs also confront increased unemployment after obtaining higher education. This counterintuitive outcome is particularly pronounced among Muslim and SC/ST groups, who, despite attaining higher education, face a disproportionately elevated risk of unemployment.

In summary, these socio-cultural determinants—caste, religion, geographic locality (rural or urban), and gender—persist in their influence, demonstrating that systemic structures and power dynamics are deeply embedded within the labor market. While it is evident that discrimination is more acute in rural areas, urban areas also reveal statistically significant disparities. Yet, education emerges as a potent site of potential disruption and upward mobility, particularly through the acquisition of higher or tertiary education, which can serve as a transformative force, albeit within the constraints of existing hierarchies. In Table 2, we present the numbers behind the above plots :

Table 2: Occupation Distribution (in %) within Fixed Education Level

				Occupation %							
Urban/Rura	Male/Female	Education	Demography	Not Working	Professional/Technical/Managerial	Clerical	Sales	Services/Household/Domestic	Agricultural	Skilled/Unskilled Manual	Other
Urban	Male	Higher	General/Upper Caste	20.445135	29.0631470	5.5124224	15.0997888	10.1966874	3.0797101	11.3612836	4.3478261
			Muslim	21.996303	20.8872458	4.8059150	21.2569316	10.1663586	3.3271719	13.8632163	3.9686577
			SC/ST	26.791277	20.0415369	6.0228453	14.5379024	11.5264798	3.0114226	14.1225337	3.9460021
Urban	Male	Primary, Secondary	General/Upper Caste	7.825639	5.1827599	2.4712545	19.4439677	12.0815887	9.3873348	37.7381157	5.8692294
			Muslim	7.502246	3.2794250	1.4824798	19.1823899	10.1976640	6.1096137	43.8903863	8.3557951
			SC/ST	9.036504	4.2195668	2.2021367	11.9118745	12.0947722	9.9701270	43.2532487	6.3100796
Urban	Male	No education, Preschool	General/Upper Caste	3.676471	1.4705882	0.0000000	13.9705882	11.7647059	15.9313725	47.7941176	5.3921569
			Muslim	2.934537	0.4514673	0.3029345	14.2212190	8.8036117	12.1896163	52.5959368	7.9006772
			SC/ST	5.191257	1.0928962	1.6393443	6.8306011	13.1147541	16.1202186	50.0000000	6.0109290
Urban	Female	Higher	General/Upper Caste	95.634573	2.5496770	0.2796211	0.3672635	0.5091667	0.0876424	0.4006511	0.1711114
			Muslim	96.804062	1.8518519	0.1698729	0.4181601	0.2090800	0.0208688	0.3285544	0.2090800
			SC/ST	95.667995	2.3912667	0.3119043	0.2252643	0.5721686	0.1386742	0.4505785	0.2777483
Urban	Female	Primary, Secondary	General/Upper Caste	95.932231	0.3830545	0.0671249	0.6578241	0.7840189	0.4806143	1.4304730	0.2577596
			Muslim	97.5983174	0.1480855	0.0267068	0.3525844	0.4447564	0.3084811	0.9918729	0.1671888
			SC/ST	95.373397	0.2891627	0.0942922	0.4777470	1.2009309	0.6977021	1.4772442	0.3834549
Urban	Female	No education, Preschool	General/Upper Caste	94.284022	0.0296165	0.0444247	0.5479046	1.3179328	1.3623575	2.2064268	0.2078153
			Muslim	97.018447	0.0214500	0.0214500	0.2788503	0.6649507	0.5148005	1.2441012	0.2359502
			SC/ST	93.255598	0.0777454	0.0777454	0.4081633	1.4965986	1.9241983	2.3129252	0.4470360
Rural	Male	Higher	General/Upper Caste	25.980199	15.3878941	3.5179369	8.5677749	5.2716579	28.2187438	9.4842785	3.6018755
			Muslim	28.703704	21.1419753	2.9320988	11.5740741	8.7962963	12.6543210	13.8024691	3.3950617
			SC/ST	29.330608	14.4098109	2.9126214	5.9765386	6.0296372	25.9070005	12.3147675	3.1170158
Rural	Male	Primary, Secondary	General/Upper Caste	7.232822	1.9890261	1.1784512	7.3887018	5.4964388	45.3203641	23.9306647	3.4605312
			Muslim	7.646742	2.0463118	1.2385568	11.9547658	6.3274098	31.5552736	33.4679591	5.7610817
			SC/ST	7.464595	1.6325728	0.9441385	4.9370574	5.0845791	47.4626279	28.4225020	4.0519276
Rural	Male	No education, Preschool	General/Upper Caste	3.611111	0.1190476	0.1190476	2.6587302	3.0158730	65.2777778	22.6587302	2.5396825
			Muslim	3.486751	0.7114438	0.0000000	5.6898507	4.3864906	45.2844875	35.8540925	4.6763345
			SC/ST	2.895323	0.2508367	0.2227171	2.3385301	2.8953229	63.4001485	24.4617669	3.5203549
Rural	Female	Higher	General/Upper Caste	96.341411	1.6839766	0.1196735	0.1453178	0.3846647	0.7522332	0.4017609	0.1709621
			Muslim	96.860168	1.3873677	0.0730194	0.0730194	0.4016064	0.2555677	0.8397225	0.1095290
			SC/ST	95.631290	1.3380604	0.1829826	0.2516011	0.4917658	1.3151876	0.5603843	0.2287283
Rural	Female	Primary, Secondary	General/Upper Caste	95.747191	0.1718178	0.0515667	0.2582374	0.3568068	2.8597609	0.8176507	0.1869789
			Muslim	97.226686	0.1277455	0.0463290	0.1680337	0.4368072	0.9766349	0.8211645	0.1936787
			SC/ST	94.256680	0.2302270	0.0640769	0.2337906	0.4779417	3.5605431	0.9178443	0.2567672
Rural	Female	No education, Preschool	General/Upper Caste	93.754049	0.0357422	0.0156372	0.1496705	0.2457277	4.8788116	0.7863286	0.1340333
			Muslim	96.520008	0.0137278	0.0343194	0.1653250	0.3431945	1.9081612	0.7898472	0.2059167
			SC/ST	92.691645	0.0430947	0.0276848	0.1622389	0.2839181	5.6073819	0.9632934	0.2205435

Note:
Total number of districts: 477
Source: Author's computation based on NFHS-5 Data

In **Table 2**, for now, we provide these figures and will revisit the statistical significance and implications of these numbers in the next section with more sophisticated statistical tools. Some of these differences may not immediately appear significant, but deeper analysis reveals important insights. For instance, over 95% of highly educated rural women are unemployed, leading to a non-significant chi-square value. However, when the 'Not working' category is excluded, significant differences in employment distribution among general, Muslim, and SC/ST groups emerge, with a p-value of 0.007. This underscores the complexities and challenges in making precise judgments for each group. Finally, it is crucial to highlight that the single most striking and statistically significant difference emerges in relation to gender. Women who have attained higher education remain largely unemployed, revealing a profound disjunction. While Figure 1 shows that men and women pursue further education almost equally across demographic groups, their employment statuses diverge markedly. For Muslim and SC/ST communities, barriers to education were more significant than barriers to occupation within same education level. In this context, it is important to recall that geographical disparity is one of the prominent sources of variability, a factor that could not be accounted for in the above calculation. To understand this further and to explore other socio-financial parameters and the broader implications of educational attainment, we have performed additional analyses on various socio-financial indicators.

Generalized regressions (global): The following section presents the Empirical Models, aimed at estimating the relationship between an outcome variable of interest (Y) and predictors (X), with a focus on the effect of education in conjunction with demographic factors. The source of dependence may stem from demographic factors alone or may not show significance when controlled for education. This combination of variations leads to the development of two types of logistic regression models:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\text{logit}\left(\text{SocioEconomic\_Indicator}_{i,r}\right) = \beta_0 + \beta_1 \cdot \text{Ed1\_Inv} + \beta_2 \cdot \text{wealth} + \sum_k \beta_k \cdot (\text{demographic\_factor}_k) + \epsilon$$

vs

$$\text{logit}\left(\text{SocioEconomic\_Indicator}_{i,r}\right) = \beta_0 + \beta_1 \cdot \text{wealth} + \sum_k \beta_k \cdot (\text{demographic\_factor}_k) + \epsilon$$

$$here\ r \in Rural, Urban,$$
$$and\ k \in \{Muslim, SC/ST, Female, General/Uppercaste\}$$

EdI_Inv = 1 - Educational Participation Index

where:

logit(p) is the log-odds of the probability $p = P(SocioEconomic\_Indicator_{i,r} = 1 \mid demographic\_factor_k\ along\ with\ other\ controlls)$

$\beta 0$ is the intercept of the model.

$\beta 1, \beta 2, \ldots, \beta k$ are the coefficients representing the effect size of each predictor (including education, wealth, and demographic factors).

ϵ represents the error term, which in the context of logistic regression, follows a binomial distribution.

The following tables summarize the results:

Table 3A: SocioEconomic Indicator: Has bank account

	Without Education Control		With Education Control	
Term	Rural	Urban	Rural	Urban
(Intercept)	2.776 (0.02)***	1.708 (0.034)***	2.729 (0.02)***	1.694 (0.034)***
WealthIndex	0.275 (0.006)***	0.354 (0.008)***	0.256 (0.006)***	0.345 (0.008)***
muslim	-0.179 (0.019)***	-0.144 (0.022)***	-0.168 (0.019)***	-0.134 (0.022)***
sc_st	-0.12 (0.015)***	0.024 (0.022)	-0.114 (0.015)***	0.028 (0.022)
female	0.003 (0.013)	0.027 (0.017)	0.033 (0.014)*	0.037 (0.017)*
EdI_inv			0.227 (0.02)***	0.081 (0.025)**

We find that participation in education positively influences the likelihood of financial inclusion, specifically in terms of having a bank account. However, the impact is considerably higher in rural areas, where the odds increase by 25.5% compared to an increase of only 8.4% in urban areas. Despite the known lower learning outcomes in rural schools, education significantly enhances financial inclusion in these areas. However, when controlled for education (i.e., within the same education groups), we do not observe any significant change in participation for Muslim or SC/ST individuals, while we see a notable change for general/upper caste women (Please note that the percentages need to be converted to exp(x) of the coefficients).

Table3B : SocioEconomic Indicator: Health Insurance (from Pvt Employer / From Government)

	Govt Insurance		Pvt Insurance (from Employer)	
Term	Without Education Control	With Education Control	Without Education Control	With Education Control
(Intercept)	-1.686 (0.006)***	-1.636 (0.006)***	-4.178 (0.048)***	-4.206 (0.049)***
WealthIndex	0.195 (0.001)***	0.227 (0.002)***	0.265 (0.01)***	0.223 (0.011)***
EdI_inv		-0.305 (0.006)***		0.314 (0.025)***
muslim	-0.454 (0.007)***	-0.483 (0.007)***	-0.649 (0.028)***	-0.598 (0.028)***
sc_st	0.276 (0.005)***	0.266 (0.005)***	-0.021 (0.022)	0.001 (0.022)
female	0.021 (0.004)***	-0.021 (0.004)***	-0.008 (0.017)	0.025 (0.017)

Note: Calculation of Pvt Insurance (from Employer) for Urban Population only

In Table 3B, a pivotal observation emerges: the coefficients for SC/ST and general/upper caste groups exhibit opposing signs when considering government versus private insurance provided by employers. This pattern reveals that SC/ST individuals are disproportionately excluded from employer-provided insurance, whereas general/upper caste individuals actively avoid government insurance. Crucially, this disparity remains pronounced even when controlling educational attainment, underscoring the persistent structural inequities at play.

Table 3C: SocioEconomic Indicator: White Collar Job (Urban)

	Comparing Caste and Religious Groups					
	Male Population		Female Population		Comparing Male and Female	
Term	Without Education Control	With Education Control	Without Education Control	With Education Control	Without Education Control	With Education Control
(Intercept)	-0.908 (0.021)***	-1.635 (0.045)***	-4.187 (0.03)***	-5.329 (0.074)***	-1.097 (0.016)***	-1.935 (0.035)***
EdI_inv		1.007 (0.052)***		1.547 (0.084)***		1.25 (0.043)***
muslim	-0.342 (0.044)***	-0.146 (0.046)**	-0.815 (0.081)***	-0.574 (0.083)***		
sc_st	-0.607 (0.044)***	-0.5 (0.045)***	-0.401 (0.064)***	-0.236 (0.065)***		
female					-3.296 (0.03)***	-3.294 (0.03)***

Note: Calculations were for Urban Population only

Table 3D:SocioEconomic Indicator: Employed (Rural)

Term	Comparing Caste and Religious Groups				Comparing Male and Female	
	Male Population		Female Population			
	Without Education Control	With Education Control	Without Education Control	With Education Control	Without Education Control	With Education Control
(Intercept)	1.423 (0.015)***	4.46 (0.045)***	-3.041 (0.011)***	-2.74 (0.013)***	1.474 (0.011)***	2.416 (0.017)***
EdI_inv		-4.209 (0.051)***		-0.748 (0.022)***		-1.553 (0.02)***
muslim	0.083 (0.036)*	-0.34 (0.042)***	-0.509 (0.029)***	-0.583 (0.029)***		
sc_st	0.121 (0.025)***	-0.194 (0.029)***	0.249 (0.016)***	0.184 (0.016)***		
female					-4.477 (0.014)***	-4.939 (0.017)***

Note: Calculations were for Rural Population only

In Tables 3C and 3D, a stark revelation emerges regarding occupational attainment across different communities. Both in rural employment and urban white-collar jobs, even after controlling for education, SC/ST and Muslim communities lag behind their similarly educated general/upper caste counterparts. SC/ST men are approximately 39% less likely to secure a white-collar job (exp(-0.5) ≈ 0.607), and Muslim men are about 14% less likely (exp(-0.146) ≈ 0.864). Similarly, SC/ST women and Muslim women are significantly disadvantaged compared to general/upper caste women, with probabilities 33% (exp(-0.40) ≈ 0.670) and 56% (exp(-0.815) ≈ 0.443) lower, respectively. In terms of overall employment, SC/ST and Muslim men are 18% (exp(-0.194) ≈ 0.824) and 29% (exp(-0.34) ≈ 0.711) less likely to be employed than their equally educated general caste counterparts. Furthermore, it appears that higher education may paradoxically disadvantage SC/ST men in rural areas, as the sign of the coefficient shifts from positive to negative when controlled for education. Muslim women in rural areas face severe employment barriers, being 44% less likely to have any kind of job compared to equally educated general/upper caste women (exp(-0.583) ≈ 0.558). Lastly, the gender disparity is the most pronounced: women are 96% less likely to hold white-collar jobs (exp(-3.294) ≈ 0.037), and in terms of rural employment, they are 99% less likely (exp(-4.939) ≈ 0.007). These findings underscore the persistent and pervasive structural inequities faced by these communities.

Table 3E:SocioEconomic Indicator: Uses Mobile Bank (Rural)

Term	Comparing Caste and Religious Groups				Comparing Male and Female	
	Male Population		Female Population			
	Without Education Control	With Education Control	Without Education Control	With Education Control	Without Education Control	With Education Control
(Intercept)	-1.411 (0.015)***	-2.762 (0.032)***	-4.301 (0.02)***	-5.273 (0.035)***	-1.575 (0.012)***	-2.74 (0.023)***
EdI_inv		2.083 (0.038)***		1.7 (0.042)***		1.933 (0.028)***
muslim	-0.218 (0.037)***	0.018 (0.039)	-0.246 (0.048)***	-0.08 (0.048)		
sc_st	-0.438 (0.027)***	-0.276 (0.028)***	-0.263 (0.034)***	-0.11 (0.035)**		
female					-2.841 (0.019)***	-2.731 (0.019)***

Note: Calculations were for Rural Population only

Lastly, in Table 3D, we investigated the impact of education on digital banking usage among various caste and religious groups, as well as among women. Contrary to the patterns observed in Table 3A, we find that mobile banking usage increases for both Muslim and SC/ST men, shifting from 20% lower participation (exp(-0.218) ≈ 0.804) to 1.8% higher participation (exp(0.018) ≈ 1.018) for Muslim men, and from 35.5% lower participation (exp(-0.438) ≈ 0.645) to 24.1% lower participation (exp(-0.276) ≈ 0.759) for SC/ST men i.e an 11% increase. Among rural women, with educational control, the coefficient for Muslim women rises from 22.1% lower participation (exp(-0.246) ≈ 0.782) to 7.7% lower participation (exp(-0.08) ≈ 0.923) i.e. a 15% increase , and for SC/ST women, it increases from 23.1% lower participation (exp(-0.263) ≈ 0.769) to 10.4% lower participation (exp(-0.11) ≈ 0.895) i.e. a 135 increase. Unlike in Table A, where physical bank account proportions were discussed, women experience a significant disparity in digital banking, reflected by 93.5% lower participation (exp(-2.731) ≈ 0.065) compared to men, instead of the 25.5% higher participation (exp(0.227) ≈ 1.255) noted for physical banking. This shift indicates that the discrimination faced by SC/ST and Muslim communities is rooted in institutional barriers within physical banking. Outside the scope of physical banking, their educational attainment markedly improves their engagement with financial services, highlighting a profound divergence in access and inclusion.

Table 3F : SocioEconomic Indicator: Can read SMS (Rural, Women)

Term	Without EdI_inv	With EdI_inv
(Intercept)	-2.43 (0.008)***	-3.548 (0.015)***
EdI_inv		1.953 (0.018)***
muslim	-0.19 (0.02)***	-0.001 (0.02)
sc_st	-0.24 (0.014)***	-0.075 (0.015)***

Table 3F: SocioEconomic Indicator: Smokes at Home (Non-Muslim Men)

Term	Without Ed1_inv	With Ed1_inv
(Intercept)	-0.24 (0.006)***	0.909 (0.008)***
Ed1_inv		-2.861 (0.007)***
sc_st	0.074 (0.005)***	0.004 (0.006)
WealthIndex	-0.238 (0.002)***	-0.071 (0.002)***

In Table 3F, we observe that the socio-economic indicator "Can read SMS" among rural Muslim women shows a substantial improvement when schooling years are controlled for. Specifically, the coefficient shifts from 17% lower participation (exp(-0.19) ≈ 0.827) to 0.1% lower participation (exp(-0.001) ≈ 0.999), indicating a 16.9% increase in participation. This indicates that schooling years significantly improve the ability of rural Muslim women to read text messages, which is crucial in the digital age. This improvement suggests that these women do not lack awareness but are subject to active discrimination. Similarly, in the same table, we see that the socio-economic indicator "Smokes at Home" for non-Muslim men also shows a marked improvement for SC/ST men when schooling years are accounted for. The coefficient for SC/ST changes from 7.7% higher participation (exp(0.074) ≈ 1.077) to 0.4% higher participation (exp(0.004) ≈ 1.004), indicating a 7.3% decrease in participation. This highlights that controlling for schooling years significantly reduces smoking at home among SC/ST men, indicating that education plays a vital role in mitigating this behavior. These findings underscore that the disparities observed are not due to a lack of awareness but rather reflect underlying discrimination. Educational attainment helps bridge these gaps, reinforcing the need for targeted educational interventions

In summary, our analysis is evidence, and it lays bare the deeply entrenched and multifaceted discrimination faced by Indian SC/ST and Muslim communities , which persists despite their educational attainment. Even with comparable pre-market endowments, such as years of schooling, these groups face formidable barriers in occupational attainment and institutionalized discrimination in physical banking. The stark disparity in returns to schooling years—where SC/ST and Muslim individuals derive significantly fewer benefits than their general/upper caste counterparts—highlights the pervasive structural inequities that education alone cannot hope to dismantle. Furthermore, their proficiency in digital banking and literacy in reading SMS messages indicates that the issue is not one of awareness but rather of systemic and deliberate discrimination. These findings make it abundantly clear that targeted interventions are necessary to address these deeply rooted and institutionalized barriers. This might be a significant barrier to promoting education in these communities.

3.5 Wealth and Social Class Transitions: A Markov Chain Analysis

Figure 29 Educational Participation Deficit Index vs Occupation Hierarchy

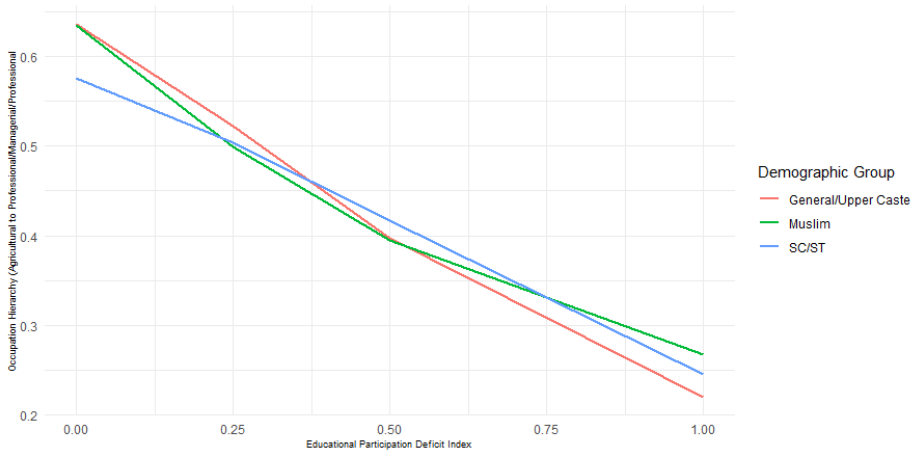


Figure 30 Wealth Index vs Household Level Educational Participation Deficit Index

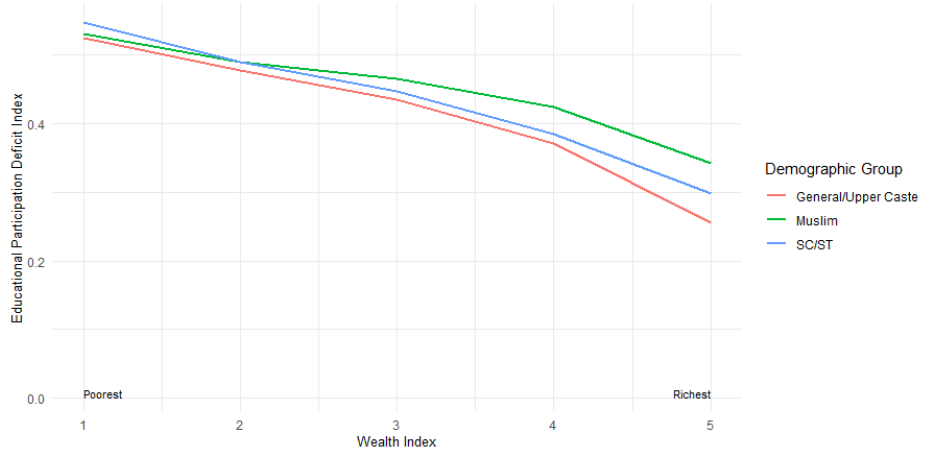


Figure 31 Household Level Educational Participation Deficit Index vs Wealth Index

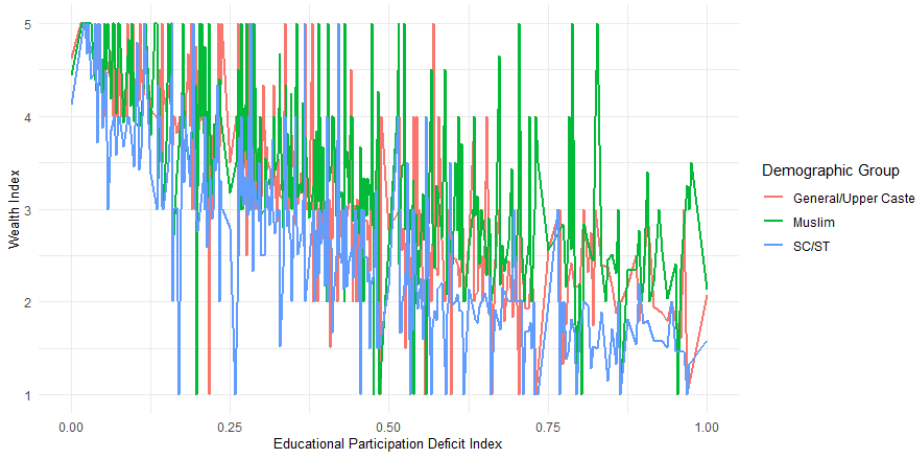
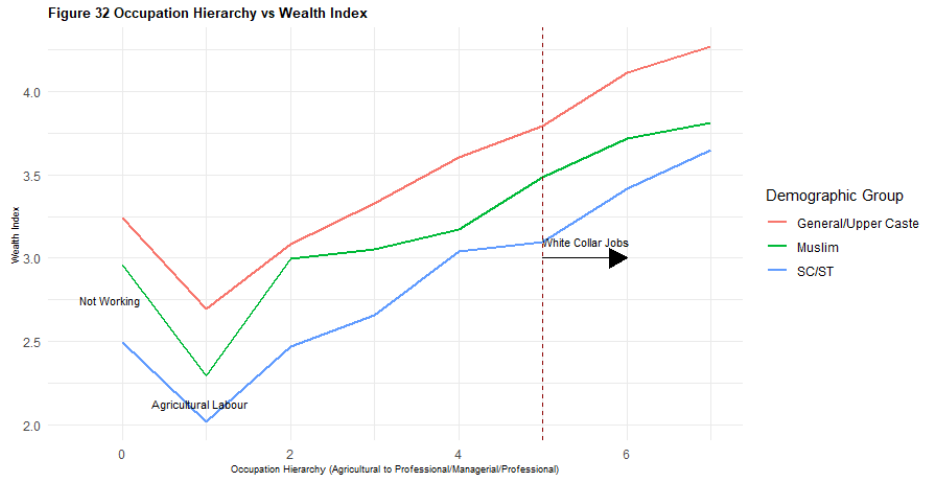


Figure 32 Occupation Hierarchy vs Wealth Index



In this section, we aim to consolidate the figures and numbers discussed so far and summarize the discussion to answer the ultimate question of transitioning from a poorer socio-economic stratum to a richer class (or vice versa). This transition is the culmination of the inequities we have observed, sequenced one after another, and situated at the intersection of these inequalities: the multifaceted barriers to educational participation shaped by wealth, demography, and geography; the differentiated entry barriers into the labor market when comparing demographic groups with equivalent educational backgrounds; and the unequal patterns of wealth accumulation from identical occupations across these groups.

Markov Chain Analysis : As discussed in Section-2 Markov chain Analysis is a statistical method used to model and analyze the mobility between different states (such as wealth levels and social classes) . It aids in understanding the long-term probability and the impact of factors like education on these transitions. By applying the mathematics of axiomatic probability, we can discern who is more likely to remain in the poorest state and who shows the most promise (or risk) of moving to adjacent states, whether upwards or downwards. We begin with a straightforward question: For an 18 to 25-year-old individual, what's the probability of starting in Wealth Category W_i and moving to Wealth Category W_j? If we had panel data tracking the same individuals over an extended period, this would be a simple matter of calculating the proportions within the aggregate sets of individuals with varying accomplishments. However, such extensive data is rarely available, particularly in a country like India, with its 1.3 billion people and immense socio-economic and geographic diversity. So, we break down the question into manageable parts: a) What is the probability of an individual from Wealth Group W_i achieving Education Level E_i b) What's the probability of an individual with Education Level E_i obtaining Occupation O_i? c) What's the probability of an individual in Occupation O_i accumulating Final Wealth W_j^f?

For example,

For general/upper-caste group

a) P (E_i | W = 'middle') = % of 18-25 years old (excluding older age groups to reflect on the current state of education barrier) with education E_i and Wealth index = 'middle'.

From our data, the probabilities stand at:

P (E_i= 6 years of education or less | W = 'middle') = 0.0929476 …(1)

P (E_i= secondary or incomplete secondary |W = 'middle') = 0.4229990 …(2)

P (E_i = completed/pursuing higher |W = 'middle') = 0.4425967 …(3)

b) P ($Occupation = O_k | E_i$) for i = 1, 2, 3 can be calculated in the same manner from the % of 18+ years old (we change the subset to all employable age groups)

The combined probability P ($Occupation = O_k$ |W = 'middle'), for general/upper-caste group will thus be :

$$\sum_{i=1}^{3} P(\text{Occupation} = O_k \mid E_i) \cdot P(E_i \mid W = \text{'middle'}) \quad (*)$$

From our data, the probabilities stand at:

P (O_i= Not working | W = 'middle') = 0.1435975 …(5)

P (O_i= Agricultural labour |W = 'middle') = 0.2778188 …(6)

P (O_i = Domestic/skilled/unskilled manual labour |W = 'middle') = 0.2984799 …(7)

P (O_i = Professional/clerical/sales |W = 'middle') = 0.2386471 …(8)

c) P (Final wealth = $W_j^f | Occupation = O_k$) for j = 1, 2, 3, 4, 5 can be calculated from the % of 18+ years old with occupation O_k and Household Wealth Index W_j^f

$$\sum_{k=1}^{4} P\left(\text{Final wealth} = W_j^f \mid \text{Occupation} = O_k\right) \cdot P(\text{Occupation} = O_k \mid W = \text{'middle'}) \quad (**)$$

$$= \sum_{k=1}^{4} P\left(\text{Final wealth} = W_j^f \mid \text{Occupation} = O_k\right) \cdot \sum_{i=1}^{3} P(\text{Occupation} = O_k \mid E_i) \cdot P(E_i \mid W = \text{'middle'}) \quad (***)$$

by inserting (*) in (**)

= P (Final wealth = W_j^f |W = 'middle')

As a general expression : P_{w_j,w_i} = P (Final wealth $= W_j^f$ |initial wealth = W_i)

$$= \sum_{k=1}^{4} P\left(\text{Final wealth} = W_j^f \mid \text{Occupation} = O_k\right) \cdot \sum_{i=1}^{3} P(\text{Occupation} = O_k \mid E_i) \cdot P(E_i \mid W = W_i) \qquad (****)$$

We will have a set of 5x5 = 25 such probability expressions which can be set as a 5X5 Markov Transition Matrix.

- **From our data , we have three Markov Chain Transition Matrices for General/Upper caste population , Muslim population and SC/ST population :**

$P_{\text{general}} =$

	poorest	poor	middle	rich	richest
poorest	0.11657082	0.1848893	0.2140796	0.2237445	0.22450269
poor	0.10836156	0.1759664	0.2083651	0.2254071	0.23836276
middle	0.10266864	0.1703416	0.2054932	0.2285444	0.25149548
rich	0.09718815	0.1649514	0.2026531	0.2313935	0.26394187
richest	0.09057321	0.1586793	0.1996417	0.2355803	0.28032367

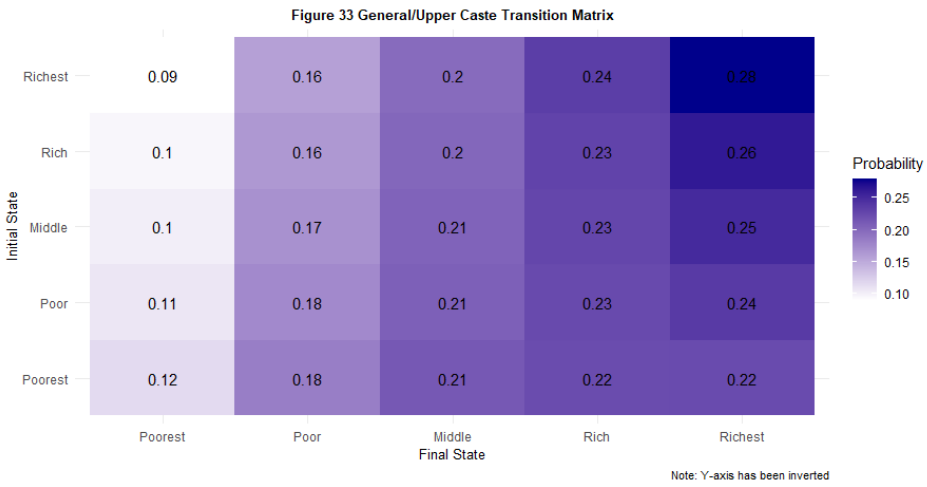
$P_{\text{muslim}} =$

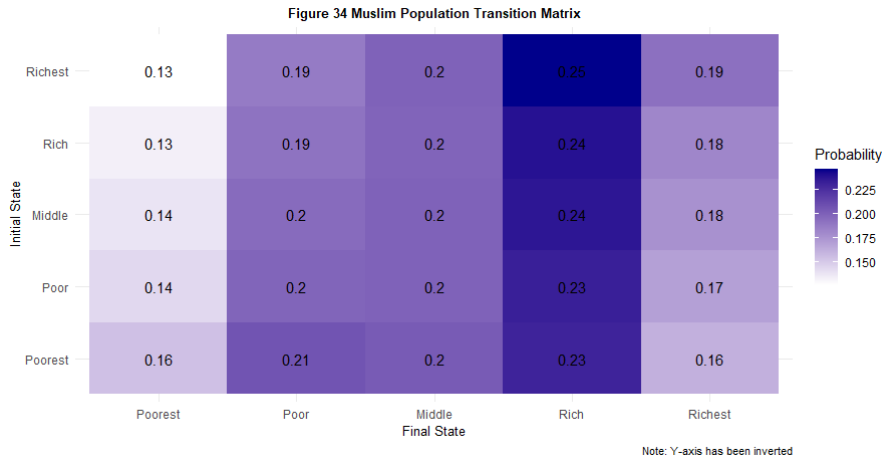
	poorest	poor	middle	rich	richest
poorest	0.15537206	0.2076527	0.2041608	0.2319291	0.16306242
poor	0.14469532	0.1991030	0.2003966	0.2339210	0.17062030
middle	0.13949065	0.1959407	0.2004520	0.2380926	0.17768783
rich	0.13437907	0.1922238	0.1994234	0.2402816	0.18263774
richest	0.12738155	0.1883174	0.2002573	0.2470518	0.19341050

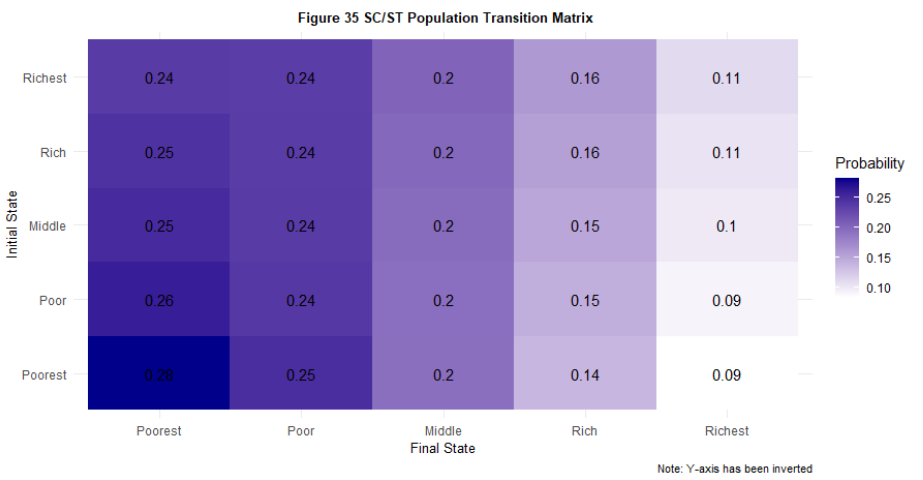
$P_{\text{SC/ST}} =$

	poorest	poor	middle	rich	richest
poorest	0.28340486	0.2496726	0.1956184	0.1386248	0.08509361
poor	0.26308712	0.2416721	0.1964580	0.1456143	0.09408568
middle	0.25240388	0.2389008	0.1988735	0.1515435	0.10094123
rich	0.24537763	0.2380838	0.2013046	0.1561066	0.10592852
richest	0.23886061	0.2373844	0.2047263	0.1622448	0.11260661

Here, the $i^{\text{th}}, j^{\text{th}}$ cell = P_{w_j,w_i} for respective demography as expressed in (****)







The Y axis has been inverted to enhance discernibility and allow for a more seamless reading of the data.

Each state transition matrix illustrates the transition probabilities of moving from one socioeconomic class to another.

In the **General/Upper Caste state transition matrix**, there is notable fluidity with significant probabilities of transitioning from lower socioeconomic classes to higher ones. For instance, individuals starting in the "Middle" state have a 23% probability of transitioning to the "Rich" state. The highest steady-state probability observed is for individuals in the "Richest" state remaining in the "Richest" state at 28%.

The **Muslim population state transition matrix** shows more restricted mobility. Although there is still some upward movement, the transition probabilities are lower compared to the General/Upper Caste group. Notably, individuals in the "Poor" state have a 24% chance of transitioning to the "Middle" state. The highest transition probability observed is for individuals in the "Middle" state remaining in the "Middle" state at 22%.

The **SC/ST population state transition matrix** highlights the challenges faced by this group. The transition probability of moving from the "Poorest" state to the "Poor" state is the highest among the three groups, reflecting significant barriers to upward mobility. The highest absorbing state probability observed is for individuals in the "Poorest" state remaining in the "Poorest" state at 28%; moreover, there is a troubling 24% probability of individuals in the "Richest" state sliding down to the "Poorest" state.

Thus, we observe distinct trajectories: clear upward mobility, constrained mobility, and disconcerting downward mobility and this dynamic is shaped by educational participation barrier, varying labor market entry opportunities, and pervasive pay gaps within the present structure. This sums up our exploratory and empirical analysis, highlighting not only disparities but also the troubling pattern observed in section 3.3, where higher educated individuals across all caste and gender categories exhibit a higher probability of remaining unemployed. The report "State of Working India 2021 – One Year of Covid-19," prepared by the Centre for Sustainable Employment at Azim Premji University, indicated that around 230 million additional individuals fell below the national minimum wage poverty line post-COVID-19. While our data analysis does not focus on any specific income shock, it offers critical insights into two areas of concern: a) the increasing likelihood of higher educated individuals remaining unemployed, and b) during economic distress, as formal salaried workers transition into informal work, a particular demographic—specifically the SC/ST population— who also had the highest education participation barriers at the district level—faces the highest risk of reverting to poverty.

Steady-State Distribution and Lerman-Yitzhaki Mobility Index : The steady-state distribution of a Markov chain represents the long-term behavior of the system. It shows the proportion of time that the system will spend in each state if it is observed over a long period. (In the context of our transition matrices, it indicates the long-term probabilities of individuals being in each wealth category unless there is any targeted intervention.) It is a probability distribution that remains unchanged as the system evolves over time. Mathematically, if $\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ is the steady-state distribution and P is the transition matrix, then :

$$\begin{aligned}
\pi_1 &= \pi_1 P_{w_1,w_1} + \pi_2 P_{w_1,w_2} + \pi_3 P_{w_1,w_3} + \pi_4 P_{w_1,w_4} + \pi_5 P_{w_1,w_5} \\
\pi_2 &= \pi_1 P_{w_2,w_1} + \pi_2 P_{w_2,w_2} + \pi_3 P_{w_2,w_3} + \pi_4 P_{w_2,w_4} + \pi_5 P_{w_2,w_5} \\
\pi_3 &= \pi_1 P_{w_3,w_1} + \pi_2 P_{w_3,w_2} + \pi_3 P_{w_3,w_3} + \pi_4 P_{w_3,w_4} + \pi_5 P_{w_3,w_5} \\
\pi_4 &= \pi_1 P_{w_4,w_1} + \pi_2 P_{w_4,w_2} + \pi_3 P_{w_4,w_3} + \pi_4 P_{w_4,w_4} + \pi_5 P_{w_4,w_5} \\
\pi_5 &= \pi_1 P_{w_5,w_1} + \pi_2 P_{w_5,w_2} + \pi_3 P_{w_5,w_3} + \pi_4 P_{w_5,w_4} + \pi_5 P_{w_5,w_5}
\end{aligned}$$

with, $\pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1$ and π_i is the long-run probability that the system (the specific demographic group) be in state i.

Solving the equations for $P_{general}, P_{muslim}, P_{SC/ST}$ we get :

$\pi_{\text{general}} = (0.1085098 \quad 0.1762993 \quad 0.2124689 \quad 0.2377772 \quad 0.2646372)$
$\pi_{\text{muslim}} = (0.1488302 \quad 0.2052756 \quad 0.2100204 \quad 0.2478190 \quad 0.1876103)$
$\pi_{\text{SC/ST}} = (0.2715868 \quad 0.2525988 \quad 0.2088015 \quad 0.1589005 \quad 0.1075531)$

- the long-term probabilities for each demographic group.

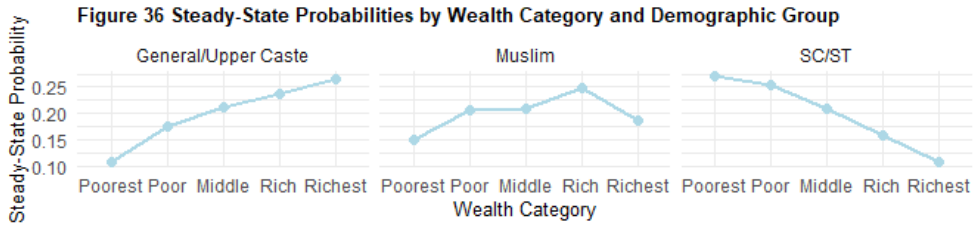


Figure 36 Steady-State Probabilities by Wealth Category and Demographic Group

Note : In the above calculation, we took the subset of male-population only so that the large unemployed female does not incorrectly influence the estimates.

Lerman-Yitzhaki Mobility Index : Finally, the Lerman and Yitzhaki Mobility Index is another metric designed to measure the extent of mobility within a system, focusing on the changes in individuals' ranks within a distribution over time. Unlike other indices (e.g., Shorrocks or Bartholomew) it can be decomposed to show both upward and downward mobility. This makes it particularly useful for understanding the directional aspects of mobility along with the magnitude . The Lerman and Yitzhaki Mobility Index is based on the concept of rank changes. It quantifies (Downward Mobility) the extent to which individuals move to lower ranks and (Upward Mobility) the extent to which individuals move to higher ranks.

$$\text{Lerman-Yitzhaki Mobility Index} = \frac{1}{N}\sum_{i=1}^{N}|d_i|$$

Where N is the total number of individuals, and $d_i = \text{Rank in Initial Period} - \text{Rank in Subsequent Period}$

$$\text{Upward Mobility U} = \frac{1}{N}\sum_{i=1}^{N} max(d_i, 0)$$
$$\text{Downward Mobility D} = \frac{1}{N}\sum_{i=1}^{N} max(-d_i, 0)$$

Solving from our $P_{general}, P_{muslim}, P_{SC/ST}$ we get :

$U_{\text{general}} = 4.552038$, $D_{\text{general}} = 2.78135$; $U_{\text{muslim}} = 3.920217$, $D_{\text{muslim}} = 3.32867$; $\text{U}_{\text{SC/ST}} = 2.626457$, $D_{\text{SC/ST}} = 4.66269$

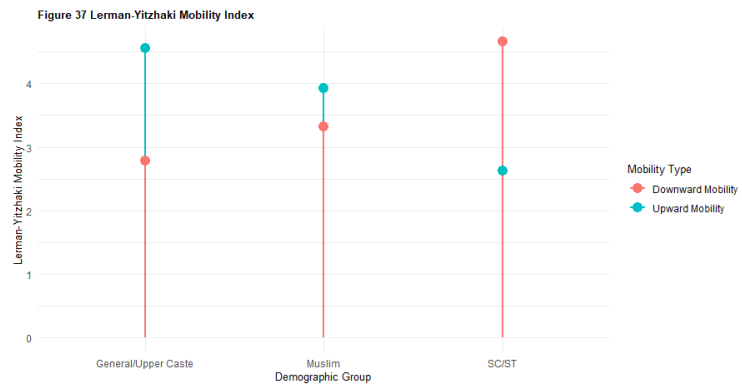


Figure 37 Lerman-Yitzhaki Mobility Index

The analysis of the Lerman-Yitzhaki Mobility Index and the steady-state probabilities by wealth category and demographic group presents a revealing picture of socio-economic mobility within India. The upward and downward mobility trends distinctly vary among the General/Upper Caste, Muslim, and SC/ST groups.

For the General/Upper Caste population, upward mobility ($U_{\text{general}} = 4.552038$) significantly surpasses downward mobility ($D_{\text{general}} = 2.78135$), suggesting a relatively fluid socio-economic landscape where individuals have a higher propensity to improve their socio-economic status. The steady-state probabilities further indicate a progressive increase from the poorest to the richest category, illustrating an overall upward trajectory in wealth accumulation. Conversely, the Muslim population exhibits a restricted mobility pattern with upward mobility ($U_{\text{muslim}} = 3.920217$) being almost same (or, slightly higher) as downward mobility ($D_{\text{muslim}} = 3.32867$). The steady-state probabilities show a peak at the middle wealth category, highlighting a plateau effect where significant upward movement beyond the middle class becomes challenging. The SC/ST population, however, faces the most significant barriers to upward mobility. With a downward mobility index ($D_{\text{SC/ST}} = 4.66269$) that far exceeds upward mobility ($\text{U}_{\text{SC/ST}} = 2.626457$), there is a pronounced risk of socio-economic regression. The steady-state probabilities for this group demonstrate a declining trend from the poorest to the richest categories, underscoring the entrenched obstacles that hinder their progress.

These findings paint a stark picture of socio-economic stratification in India, where systemic inequities manifest in differing mobility patterns across demographic groups. The data underscores the critical need for targeted interventions to address these disparities, ensuring equitable opportunities for upward mobility across all segments of the population.