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^{th}*district and the estimated* $\beta_i = (X^T W_i X)^{-1} X^T W_i Y$. So, all β_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^{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.

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

$k = 1, 2, \dots, S$ *and* *i = 1, 2,…,* 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, $\mathrm{I_i} = \frac{(Y_i - \overline{Y})}{m_2} \sum_j w_{ij}\left(Y_j - \overline{Y}\right)$ (3)

Where $\mathrm{m_2}$ 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 :

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

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

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, $\widehat{\theta_i} = \overline{Y}$ 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 :

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

$Here$ $\text{demographic factor}_k \in \{\text{Muslim}, \text{SC/ST}, \text{Female}\}$; 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= 12 . And for each of these 19 clusters we assume that the distribution of district level parameters follows latent distribution $G_1, G_2, \dots G_{12}$ 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.

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
NOT WORKING	8	Not Working

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

2.3.2 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 $= W_j^f \mid$ Occupation $= O_k \Big) \cdot \sum_{i=1}^{3} P($ Occupation $= O_k \mid E_i) \cdot$
$P(E_i \mid W = W_i)$

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 ($\text{Occupation} = O_k$ |W = ‘middle’), for general/upper-caste group will thus be :

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

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 |$\text{Occupation} = \text{O}_\text{k}$) 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:

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

by inserting (*) in (**) we get

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

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 :

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

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.

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

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

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

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

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.

2.4 Do We have Consistent Evidence of Differences Outside the Labor Market?

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:

In regression analysis when we have binary outcome variable of interest, instead of using the variable itself we use a log function, $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ on the left-hand side of the regression equation. Here, p= P(Y =1).

Our regression equation is as follows :

$$\text{logit}\big(\text{p}(\text{SocioEconomic Indicator}_{i,r})\big) = \beta_0 + \beta_1 \cdot (\text{1-E}) + \beta_2 \cdot \text{wealth} + \sum_k \beta_k \cdot (\text{demographic factor}_k) + \epsilon$$

vs

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

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

We also denote,
Ed1_Inv= 1-E = 1 - Educational Participation Index

logit(p) is the log-odds of the probability

$\text{logit}(\text{P}(\text{SocioEconomic Indicator}_{\text{i,r}} = 1 \,|\text{demographic factor}_\text{k}$ along with other controlls))

β_0 is the intercept of the model. $\beta_1, \beta_2, \ldots, \beta_\text{k}$ 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.

In this analysis, we compare two models—one with education as a control variable and one without—to determine if the observed disparities in labour market or wealth mobility extend to other dimensions such as environmental consciousness and hygiene. It is a widespread belief that socially marginalized communities are inherently prone to various malpractices. However, our findings reveal that, unlike the discrepancies in the job market, when education is controlled for, SC/ST and Muslim communities exhibit civic habits comparable to those of general or upper castes.

3. Empirical Results and Discussions

Following our deliberations on the regional disparities in participation and the nuances of within-region participation and autocorrelation, it becomes imperative to engage with what our NFHS-5 data elucidates regarding these metrics and their statistical significance. We will deploy our composite measure of Educational Participation Deficit (E), as delineated in Section 2.3.1, alongside other straightforward measures of educational participation, such as the number of years of education.

These will be used interchangeably, depending on whether the analysis necessitates a discrete or continuous measure. The forthcoming sections are structured as follows: In Section 3.1.1 and 3.1.2, we will disclose the summary results, while Section 3.2 will be devoted to the outcomes of our regression analysis. Subsequently, Section 3.3 will illuminate the findings from our Markov Chain Analysis, and ultimately, Section 3.4 will elucidate the inter-community net mobility and its relationship with educational participation. Additionally, in Section 3.5, we will present supplementary and confirmatory logistic regression results, as discussed in Section 2.4, to interrogate the implications of educational participation beyond wealth mobility.

Our analysis suggests that overall average participation within a district exerts an insignificant influence on reducing inequality (Figure 1-2). This inverse relationship fails to provide any optimism, urging a cautious approach in policy-making. Rather than merely enhancing the average condition, policies must be meticulously designed to address specific inequality issues within the districts . Along with that, intra-district inequality persists at elevated levels. Notably, we will also illustrate that the mechanisms of this inequality in educational participation are not rooted in wealth per se, as well as wealth mobility appears to be more contingent upon demographic variables and regional (district level) disparities than upon educational attainment, although educational participation and wealth-mobility are not completely unrelated. Also, we will illustrate that districts exhibit a tendency to form spatial clusters with unique cluster maps at the conclusion of section 3.4. These clusters traverse multiple administrative states, providing us a geographical reality within the clusters that is more homogenous than the state borders, which have traditionally played a pivotal role in policy-making.

3.1.1 Summary Statistics and Key Spatial Exploratory Data Analysis Results

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 .

Figure-1 District Level Participation Deficit (E)

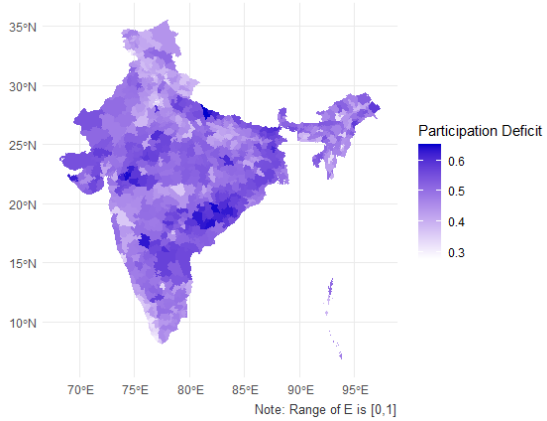
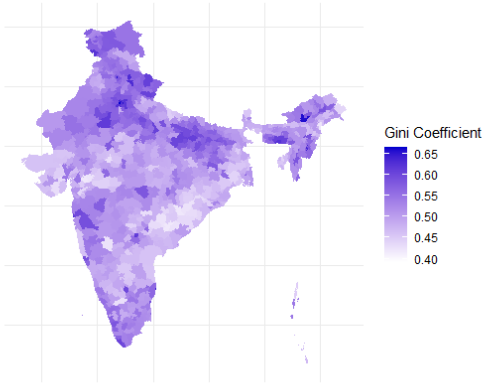
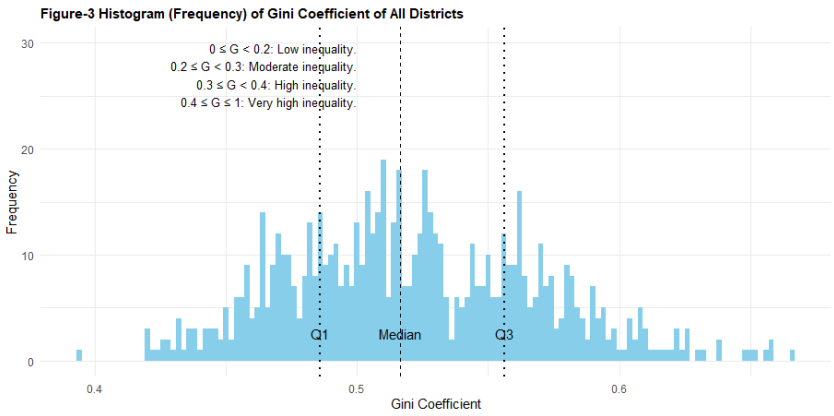


Figure-2 District-wise Gini Coefficient





Figures 1 and 2 demonstrate that regions with low participation deficit (indicated in lighter colors) exhibit greater participation inequality (shown in darker colors). The Pearson's coefficient is -0.86, and the Spearman's Rank Correlation is -0.87. Conversely, the median Gini Coefficient is 0.52, reflecting very high inequality across all districts. This suggests that demographic socio-economic groups with lower participation levels do not experience significant improvement, even as overall participation increases.

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.

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 2). 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 . 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. In Table-1, we 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

Urban/Rural	Age Group	Demographic Parameter	Demography	Sample Size	Mean (SD)	t Test Statistic	F Test Statistic
Urban	Below 12 years	Gender	female	56857	0.0256 (0.1580)	0.77	
			male	62406	0.0249 (0.1559)		
		Caste	General/Others	31031	0.0190 (0.1365)		34.896***
			Other Backward Class	50416	0.0267 (0.1611)		
			Scheduled Caste/Tribe	37816	0.0285 (0.1665)		
		Religion	Hindu/Jain	83874	0.0220 (0.1468)		127.921***

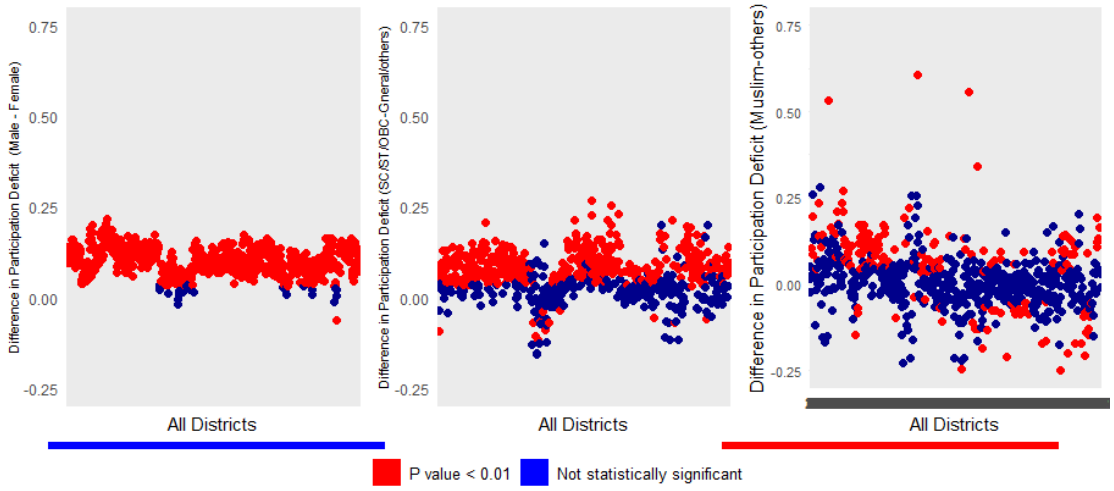
			Muslim	23017	0.0428 (0.2025)		
			Sikh	2179	0.0133 (0.1146)		
			Others	10193	0.0147 (0.1204)		
	12 to 18 years old	Gender	female	37903	0.1222 (0.2882)	-6.88***	
			male	41300	0.1367 (0.3035)		
		Caste	General/Others	21090	0.0915 (0.2536)		267.047***
			Other Backward Class	33127	0.1360 (0.3024)		
			Scheduled Caste/Tribe	24986	0.1538 (0.3179)		
		Religion	Hindu/Jain	55995	0.1110 (0.2765)		610.109***
			Muslim	14817	0.2218 (0.3673)		
			Sikh	1581	0.1013 (0.2653)		
			Others	6810	0.0902 (0.2453)		
	18+ years old	Gender	female	227164	0.5302 (0.3892)	99.359***	
			male	221589	0.4191 (0.3588)		
		Caste	General/Others	134729	0.3816 (0.3663)		6386.097***
			Other Backward Class	184729	0.5025 (0.3772)		
			Scheduled Caste/Tribe	129295	0.5350 (0.3751)		
		Religion	Hindu/Jain	335547	0.4591 (0.3791)		2466.669***
			Muslim	66356	0.5893 (0.3686)		
			Sikh	10557	0.4492 (0.3751)		
			Others	36293	0.4266 (0.3547)		
Rural	Below 12 years	Gender	female	219433	0.0379 (0.1909)	3.395***	
			male	235255	0.0360 (0.1864)		
		Caste	General/Others	70425	0.0244 (0.1542)		332.345***
			Other Backward Class	174149	0.0335 (0.1799)		
			Scheduled Caste/Tribe	210114	0.0440 (0.2051)		
		Religion	Hindu/Jain	348120	0.0344 (0.1821)		397.699***
			Muslim	48262	0.0636 (0.2440)		
			Sikh	8753	0.0128 (0.1124)		
			Others	49553	0.0333 (0.1794)		
	12 to 18 years old	Gender	female	136971	0.2093 (0.3558)	23.173***	
			male	139862	0.1789 (0.3343)		
		Caste	General/Others	44846	0.1288 (0.2917)		1370.437***
			Other Backward Class	108640	0.1848 (0.3376)		
			Scheduled Caste/Tribe	123347	0.2257 (0.3658)		
		Religion	Hindu/Jain	215825	0.1863 (0.3399)		727.232***
			Muslim	28837	0.2792 (0.3963)		
			Sikh	5852	0.1214 (0.2793)		
			Others	26319	0.1798 (0.3292)		
	18+ years old	Gender	female	684389	0.7313 (0.3447)	226.917***	
			male	638766	0.5927 (0.3578)		
		Caste	General/Others	249970	0.5722 (0.3697)		12371.23***
			Other Backward Class	509592	0.6638 (0.3589)		
			Scheduled Caste/Tribe	563593	0.7062 (0.3435)		
		Religion	Hindu/Jain	1044353	0.6646 (0.3593)		478.747***
			Muslim	109712	0.6942 (0.3542)		
			Sikh	35643	0.6240 (0.3554)		
			Others	133447	0.6499 (0.3478)		

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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 (Figure -4) . 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 4: Statistical Siginifcance Tests Within each District



In Fig -4 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 and Fig 3).

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.

3.1.2 Regional High and Low Educational Participation Zones

In this segment, our objective is to identify groups of districts with similar educational engagement, essential for our forthcoming multi-context analysis. Numerous mechanisms, likely functioning

through spatial or physical proximity, are significant (Soja, 2013). While we avoid a deep dive into spatial econometrics, examining spatial likeness and difference is crucial to understanding educational participation disparities and their connection to wealth mobility. When we draw the cluster maps, we will see the general frameworks of pooling or segregating data at the national, state (administrative level -1), or urban/rural levels do not offer effectively clearer explanations. In the subsequent analysis, we will use these new zone borders or identifiers as covariates, as is typical in statistical discussions. Theoretically, regional reality fluctuates between two primary aspects: the influence of socioeconomic, demographic, empowerment, and other factors from neighboring districts on educational outcomes and inequalities, indicating spatial dependence. Secondly, identifying which regions exhibit minimal effects from their neighbors involves understanding spatial heterogeneity, where the relationships between variables change across locations, making some areas uninfluenced by adjacent districts despite the overall spatial trends. And to understand the two we start with the concept of spatial weight . Greater weight is assigned to nearby districts and those sharing borders, with influence diminishing as distance increases. This 'weight' quantifies our baseline assumption on the impact of one area on another, helping us finally discern which areas are more or less influenced by their neighbors, even if causal channels remain unclear.

Selecting an appropriate weighting matrix has various potential options where nearest neighbors are always assigned higher weights. For our study, we define $w_{ij} = \exp\left(-\frac{|\text{distance}|}{\text{threshold}}\right)$ 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 .

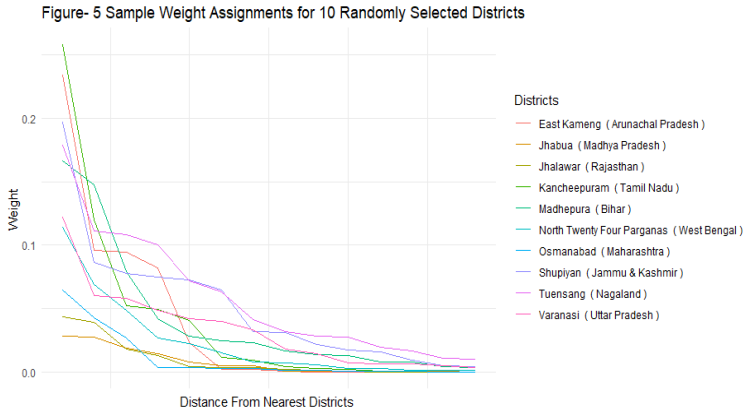


Figure 5 illustrates this rate of decay i.e. what a threshold of 0.2 means, and how the influence diminishes with increasing distance. For example, weights assigned to Jhalwar (Rajasthan), where neighboring districts are larger, sharply decay to zero approximately after the 5th district. In contrast, 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.

The Local Indicator of Spatial Autocorrelation will be used to draw the spatial clusters where in Figure 5 we discussed how the weight assignment is done.

$$\text{LISA } I_i = \frac{x_i - \bar{x}}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \sum_{j \neq i} w_{ij}\left(x_j - \bar{x}\right) \quad \text{(in section 2.1)}$$

The observations $(x_1, x_2 \ \ldots x_N)$ 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. 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.

Figure 6A Simple LISA Values For Each Districts

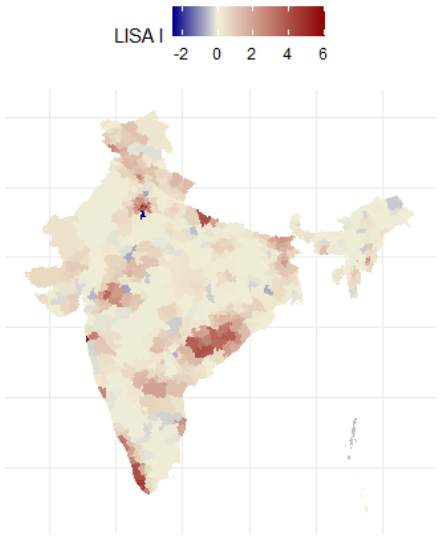
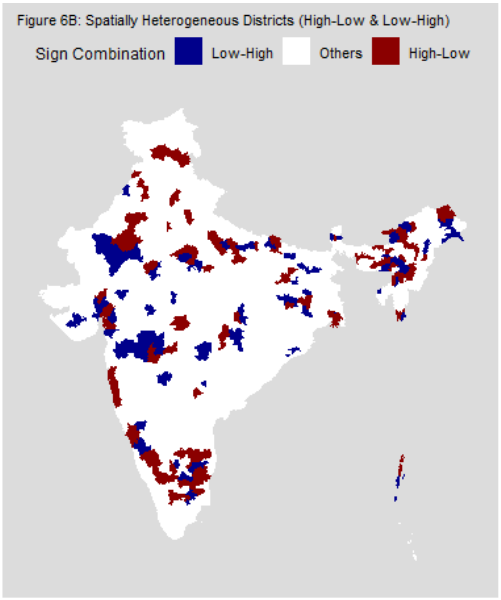
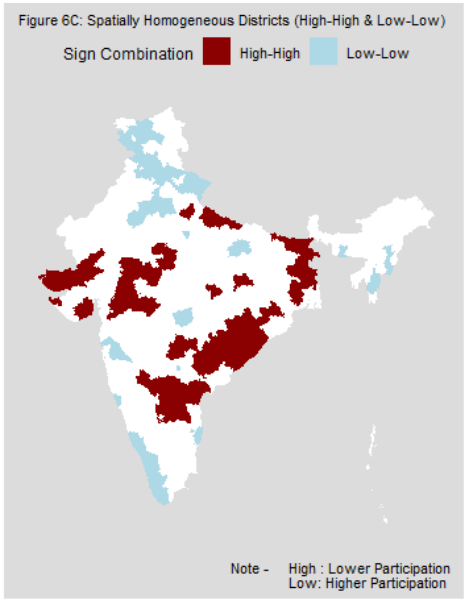


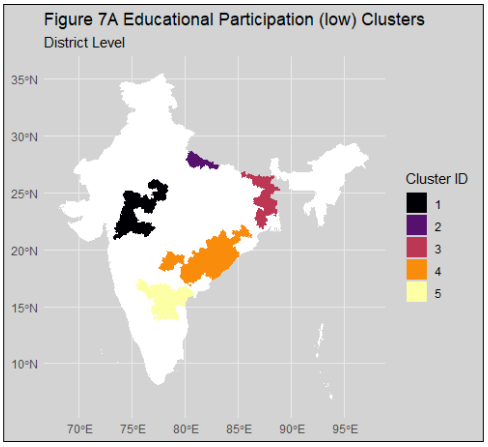
Figure 6A displays the LISA values for each district, which depend on both the magnitude of correlation and the direction of association. The red zones indicate positive LISA values. However, it is unclear which of these are positive because they are surrounded by other districts with high educational participation deficits (forming a hot spot region or cluster of districts with low educational participation) and which are positive due to being surrounded by districts with high educational participation. This distinction is clarified in Figures 6B and 6C.





Map 6C is of interest to us and we want to mark the different clusters and assign them a cluster ID which is our final goal from the Spatial Exploratory Analysis.

We identified a total of 12 spatial clusters. Below we have listed and mapped the clusters :



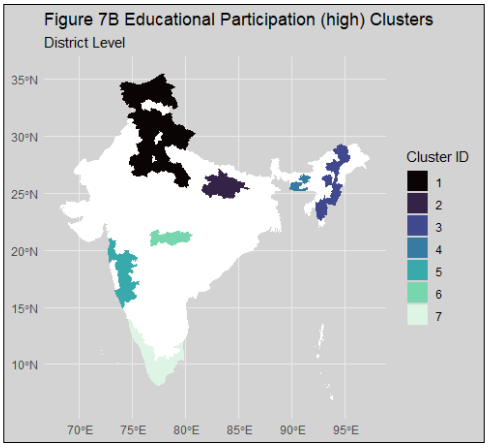


Table 2A : HIGH Educational Participation Clusters :

Cluster ID	Districts (State)	Number of Districts
1	Kupwara(Jammu & Kashmir), Badgam(Jammu & Kashmir), Leh(Ladakh), Kargil(Ladakh), Punch(Jammu & Kashmir), Rajouri(Jammu & Kashmir), Kathua(Jammu & Kashmir), Baramula(Jammu & Kashmir), Bandipore(Jammu & Kashmir), Srinagar(Jammu & Kashmir), Ganderbal(Jammu & Kashmir), Pulwama(Jammu & Kashmir), Shupiyan(Jammu & Kashmir), Anantnag(Jammu & Kashmir), Kulgam(Jammu & Kashmir), Doda(Jammu & Kashmir), Ramban(Jammu & Kashmir), Kishtwar(Jammu & Kashmir), Udhampur(Jammu & Kashmir), Reasi(Jammu & Kashmir), Jammu(Jammu & Kashmir), Samba(Jammu & Kashmir), Kangra(Himachal Pradesh), Kullu(Himachal Pradesh), Mandi(Himachal Pradesh), Hamirpur(Himachal Pradesh), Hamirpur(Uttar Pradesh), Una(Himachal Pradesh), Bilaspur(Himachal Pradesh), Solan(Himachal Pradesh), Sirmaur(Himachal Pradesh), Shimla(Himachal Pradesh), Kinnaur(Himachal Pradesh), Kapurthala(Punjab), Jalandhar(Punjab), Hoshiarpur(Punjab), Shahid Bhagat Singh Nagar(Punjab), Fatehgarh Sahib(Punjab), Ludhiana(Punjab), Patiala(Punjab), Amritsar(Punjab), Rupnagar(Punjab), Sahibzada Ajit Singh Nagar(Punjab), Sangrur(Punjab), Chandigarh(Chandigarh), Uttarkashi(Uttarakhand), Chamoli(Uttarakhand), Rudraprayag(Uttarakhand), Tehri Garhwal(Uttarakhand), Dehradun(Uttarakhand), Garhwal(Uttarakhand), Pithoragarh(Uttarakhand), Bageshwar(Uttarakhand), Almora(Uttarakhand), Champawat(Uttarakhand), Nainital(Uttarakhand), Udham Singh Nagar(Uttarakhand), Hardwar(Uttarakhand), Panchkula(Haryana), Ambala(Haryana), Yamunanagar(Haryana), Kurukshetra(Haryana), Kaithal(Haryana), Karnal(Haryana), Panipat(Haryana), Sonipat(Haryana), Jind(Haryana), Hisar(Haryana), Rohtak(Haryana), Jhajjar(Haryana), Mahendragarh(Haryana), Rewari(Haryana), Gurgaon(Haryana), Faridabad(Haryana), Palwal(Haryana), Jhunjhunun(Rajasthan), Alwar(Rajasthan), Bharatpur(Rajasthan), Dausa(Rajasthan), Jaipur(Rajasthan), Sikar(Rajasthan), Bijnor(Uttar Pradesh), Meerut(Uttar Pradesh), Baghpat(Uttar Pradesh), Gautam Buddha Nagar(Uttar Pradesh), Bulandshahr(Uttar Pradesh), Aligarh(Uttar Pradesh), Mahamaya Nagar(Uttar Pradesh), Mathura(Uttar Pradesh), Agra(Uttar Pradesh), Firozabad(Uttar Pradesh), Mainpuri(Uttar Pradesh), Farrukhabad(Uttar Pradesh), Kannauj(Uttar Pradesh), Etawah(Uttar Pradesh), Auraiya(Uttar Pradesh), Kanpur Dehat(Uttar Pradesh), Kanpur Nagar(Uttar Pradesh), Jalaun(Uttar Pradesh), Etah(Uttar Pradesh), Central(Nct Of Delhi), East(Nct Of Delhi), New Delhi(Nct Of Delhi), North(Nct Of Delhi), North East(Nct Of Delhi), North West(Nct Of Delhi), Shahdara(Nct Of Delhi), South(Nct Of Delhi), South East(Nct Of Delhi), South West(Nct Of Delhi), West(Nct Of Delhi), Bhiwani(Haryana), Charkhi Dadri(Haryana), Gurdaspur(Punjab), Pathankot(Punjab), Ghaziabad(Uttar Pradesh), Hapur(Uttar Pradesh), Muzaffarnagar(Uttar Pradesh)	118

2	Allahabad(Uttar Pradesh), Ambedkar Nagar(Uttar Pradesh), Gorakhpur(Uttar Pradesh), Deoria(Uttar Pradesh), Azamgarh(Uttar Pradesh), Mau(Uttar Pradesh), Ballia(Uttar Pradesh), Jaunpur(Uttar Pradesh), Ghazipur(Uttar Pradesh), Chandauli(Uttar Pradesh), Varanasi(Uttar Pradesh), Sant Ravidas Nagar(Uttar Pradesh), Mirzapur(Uttar Pradesh), Siwan(Bihar), Saran(Bihar), Vaishali(Bihar), Patna(Bihar), Bhojpur(Bihar), Buxer(Bihar), Kaimur (Bhabua)(Bihar), Rohtas(Bihar), Aurangabad(Bihar), Jehanabad(Bihar), Arwal(Bihar), Sultanpur(Uttar Pradesh)	25
3	Papum Pare(Arunachal Pradesh), Upper Siang(Arunachal Pradesh), Lower Subansiri(Arunachal Pradesh), Wokha(Nagaland), Dimapur(Nagaland), Phek(Nagaland), Kohima(Nagaland), Senapati(Manipur), Churachandpur(Manipur), Bishnupur(Manipur), Thoubal(Manipur), Imphal West(Manipur), Imphal East(Manipur), Ukhrul(Manipur), Chandel(Manipur), Mamit(Mizoram), Kolasib(Mizoram), Aizawl(Mizoram), Champhai(Mizoram), Serchhip(Mizoram), Lunglei(Mizoram), Lakhimpur(Assam), Dhemaji(Assam), East Siang(Arunachal Pradesh), Siang(Arunachal Pradesh), West Siang(Arunachal Pradesh), Jorhat(Assam), Karbi Anglong(Assam), Majuli(Assam)	29
4	South Garo Hills(Meghalaya), Barpeta(Assam), Kamrup(Assam), Nalbari(Assam), East Garo Hills(Meghalaya), North Garo Hills(Meghalaya), South West Garo Hills(Meghalaya), South West Khasi Hills(Meghalaya), West Garo Hills(Meghalaya)	9
5	Navsari(Gujarat), Valsad(Gujarat), Daman(Dadra & Nagar Haveli & Daman & Diu), Dadra & Nagar Haveli(Dadra & Nagar Haveli & Daman & Diu), Mumbai Suburban(Maharashtra), Mumbai(Maharashtra), Pune(Maharashtra), Ahmadnagar(Maharashtra), Satara(Maharashtra), Sindhudurg(Maharashtra), Kolhapur(Maharashtra), Sangli(Maharashtra), Belgaum(Karnataka), North Goa(Goa), South Goa(Goa), Palghar(Maharashtra), Thane(Maharashtra)	17
6	Akola(Maharashtra), Amravati(Maharashtra), Wardha(Maharashtra), Nagpur(Maharashtra), Bhandara(Maharashtra), Gondiya(Maharashtra)	6
7	Udupi(Karnataka), Dakshina Kannada(Karnataka), Kodagu(Karnataka), Kasaragod(Kerala), Kannur(Kerala), Wayanad(Kerala), Kozhikode(Kerala), Malappuram(Kerala), Palakkad(Kerala), Thrissur(Kerala), Ernakulam(Kerala), Idukki(Kerala), Kottayam(Kerala), Alappuzha(Kerala), Pathanamthitta(Kerala), Kollam(Kerala), Thiruvananthapuram(Kerala), The Nilgiris(Tamil Nadu), Dindigul(Tamil Nadu), Cuddalore(Tamil Nadu), Nagapattinam(Tamil Nadu), Thiruvarur(Tamil Nadu), Thanjavur(Tamil Nadu), Pudukkottai(Tamil Nadu), Sivaganga(Tamil Nadu), Madurai(Tamil Nadu), Theni(Tamil Nadu), Virudhunagar(Tamil Nadu), Ramanathapuram(Tamil Nadu), Thoothukkudi(Tamil Nadu), Tirunelveli(Tamil Nadu), Kanniyakumari(Tamil Nadu), Coimbatore(Tamil Nadu), Puducherry(Puducherry), Mahe(Puducherry), Karaikal(Puducherry)	36

Table 2B : LOW Educational Participation Clusters :

Cluster ID	Districts (States)	Number of Districts
1	Banswara(Rajasthan), Chittaurgarh(Rajasthan), Jhalawar(Rajasthan), Pratapgarh(Rajasthan), Sheopur(Madhya Pradesh), Shivpuri(Madhya Pradesh), Neemuch(Madhya Pradesh), Mandsaur(Madhya Pradesh), Ratlam(Madhya Pradesh), Dhar(Madhya Pradesh), Khargone (West Nimar)(Madhya Pradesh), Barwani(Madhya Pradesh), Rajgarh(Madhya Pradesh), Guna(Madhya Pradesh), Ashoknagar(Madhya Pradesh), Jhabua(Madhya Pradesh), Alirajpur(Madhya Pradesh), Khandwa (East Nimar)(Madhya Pradesh), Burhanpur(Madhya Pradesh), Dohad(Gujarat), Narmada(Gujarat), Tapi(Gujarat), Nandurbar(Maharashtra), Chhota Udaipur(Gujarat), Agar Malwa(Madhya Pradesh), Shajapur(Madhya Pradesh)	26
2	Kheri(Uttar Pradesh), Sitapur(Uttar Pradesh), Bahraich(Uttar Pradesh), Shrawasti(Uttar Pradesh), Balrampur(Uttar Pradesh), Siddharthnagar(Uttar Pradesh)	6
3	Sitamarhi(Bihar), Madhubani(Bihar), Supaul(Bihar), Araria(Bihar), Kishanganj(Bihar), Purnia(Bihar), Katihar(Bihar), Madhepura(Bihar), Saharsa(Bihar), Uttar Dinajpur(West Bengal), Dakshin Dinajpur(West Bengal), Maldah(West Bengal), Murshidabad(West Bengal), Birbhum(West Bengal), Nadia(West Bengal), Bankura(West Bengal), Paschim Medinipur(West Bengal), Deoghar(Jharkhand), Godda(Jharkhand), Sahibganj(Jharkhand), Pakur(Jharkhand), Dumka(Jharkhand), Jamtara(Jharkhand), Purba Barddhaman(West Bengal)	24

4	Bargarh(Odisha), Debagarh(Odisha), Kendujhar(Odisha), Anugul(Odisha), Ganjam(Odisha), Gajapati(Odisha), Kandhamal(Odisha), Baudh(Odisha), Subarnapur(Odisha), Balangir(Odisha), Nuapada(Odisha), Kalahandi(Odisha), Rayagada(Odisha), Nabarangapur(Odisha), Koraput(Odisha), Malkangiri(Odisha), Narayanpur(Chhattisgarh), Bijapur(Chhattisgarh), Srikakulam(Andhra Pradesh), Vizianagaram(Andhra Pradesh), Visakhapatnam(Andhra Pradesh), Bastar(Chhattisgarh), Dantewada(Chhattisgarh), Gariyaband(Chhattisgarh), Kodagaon(Chhattisgarh), Sukma(Chhattisgarh), Adilabad(Telangana), Bhadradri Kothagudem(Telangana), Jagitial(Telangana), Jayashankar Bhupalapally(Telangana), Kamareddy(Telangana), Khammam(Telangana), Komaram Bheem Asifabad(Telangana), Mahabubabad(Telangana), Mancherial(Telangana), Nirmal(Telangana), Nizamabad(Telangana), Rajanna Sircilla(Telangana), Warangal Rural(Telangana)	39
5	Bijapur(Karnataka), Guntur(Andhra Pradesh), Prakasam(Andhra Pradesh), Y.S.R.(Andhra Pradesh), Kurnool(Andhra Pradesh), Anantapur(Andhra Pradesh), Raichur(Karnataka), Yadgir(Karnataka), Jogulamba Gadwal(Telangana), Mahabubnagar(Telangana), Nagarkurnool(Telangana), Wanaparthy(Telangana)	12

It is imperative to recognize the significant role these clusters will play in our forthcoming analysis. Initially, they will be pivotal in our Empirical Bayesian Analysis. Subsequently, they will serve as a reference point when examining the interplay between inter-community wealth mobility and education. Notably, these clusters transcend demographic distinctions; they are defined by overall educational participation rather than being specific to any particular demographic entity.

3.2 District Level Repeated Mixed-Effect Regressions and Bayesian Approximation.

We will reiterate the formula for our within district modelling and state the results.

Stage – I

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

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

$$;\ p = 4,5$$

$\text{Number of Years Education}_{ijd}$ = 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.

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

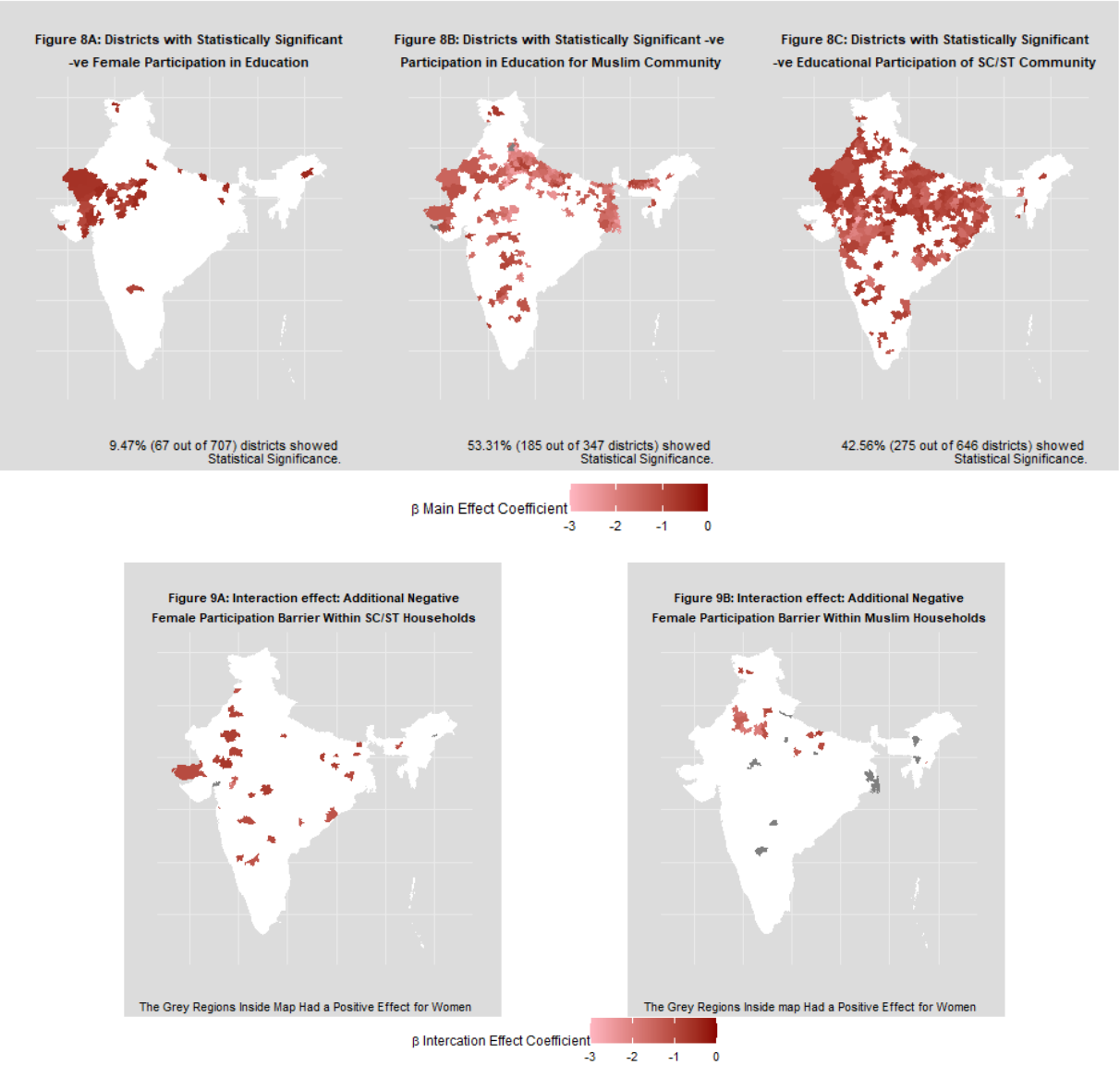
The above equation has two other variations depending on the 'Rank' constraint where adequate data on a specific demographic group is not available and those have been delineated in section 2.2 and we will not repeat that again here.

Stage – II

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

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

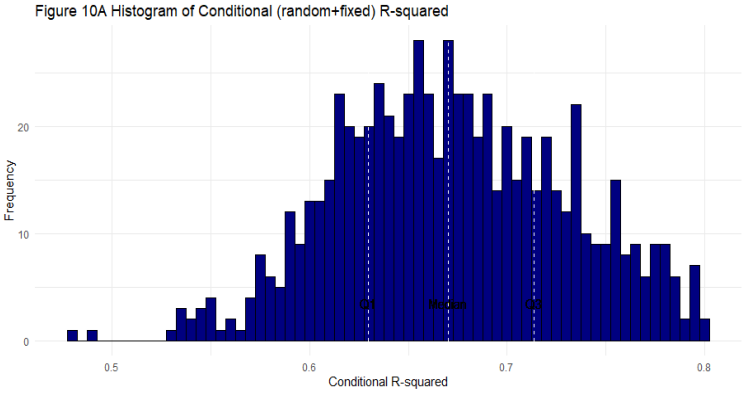
There exists a multitude of studies comparing the efficiency of Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR), with a consistent emphasis on the superior efficiency of GWR. However, a global OLS and a regional-level GWR, even in the absence of spatial autocorrelation, are inherently incomparable due to their differing scales; OLS operates on pooled data, whereas GWR is applied to more granular, group-level data. A more appropriate comparison would be between repeated OLS and GWR. Naturally, if spatial autocorrelation is present, GWR is preferable. Nonetheless, GWR is both time and memory intensive. To address this, we devised a method involving multiple hierarchical models. While not full or hierarchical Bayes, even this Empirical Bayes approach yields satisfactory predictive power, evidenced by a median R-squared of 0.8, which is notably high. In Figures 8A to 8C, we display the main effect results, while Figures 8E and 8F illustrate the interaction effects. Additionally, Figure 9 presents the R-squared values for each district, along with their distribution.

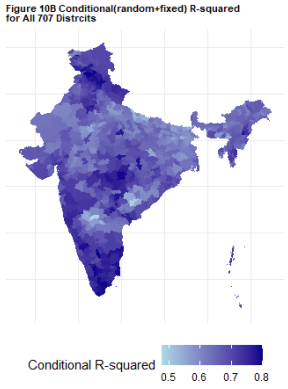


Figures 8A-8C display the regression summary results (magnitude and p-values) of the 707 regressions we conducted. In the maps, the main-effect coefficients ($\beta_{\text{female,d}}$, $\beta_{\text{SC/ST,d}}$ and $\beta_{\text{muslim,d}}$) for districts d= 1, 2… 707 are provided for each district where the coefficients were statistically significant. If a

particular β estimate is absent for a district in the maps, it indicates that no significant negative 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 Figures 10A and 10B

Another crucial point to acknowledge is our exploration of interaction effects, representing the non-linear dynamics beyond the additive impacts of being female and part of Muslim/SC/ST communities. This analysis challenges the widespread belief that these households are particularly regressive toward women, at least concerning barriers to educational participation in India, based on our 2019-21 data. Additionally, we observe a shift in the number of significant districts compared to Figure 4. Previously, many districts showed significance for women and fewer for Muslims and SC/ST communities, but this did not address the combined channels of inequality. Our analysis, incorporating household-level random effects, reveals a considerable shift in the results. Lastly, it is important to note that the districts in Map 6B do not require weighted regression or borrowing from neighboring areas, as their LISA values show no correlation with adjacent regions. Therefore, shrinking their estimates would not be effective.





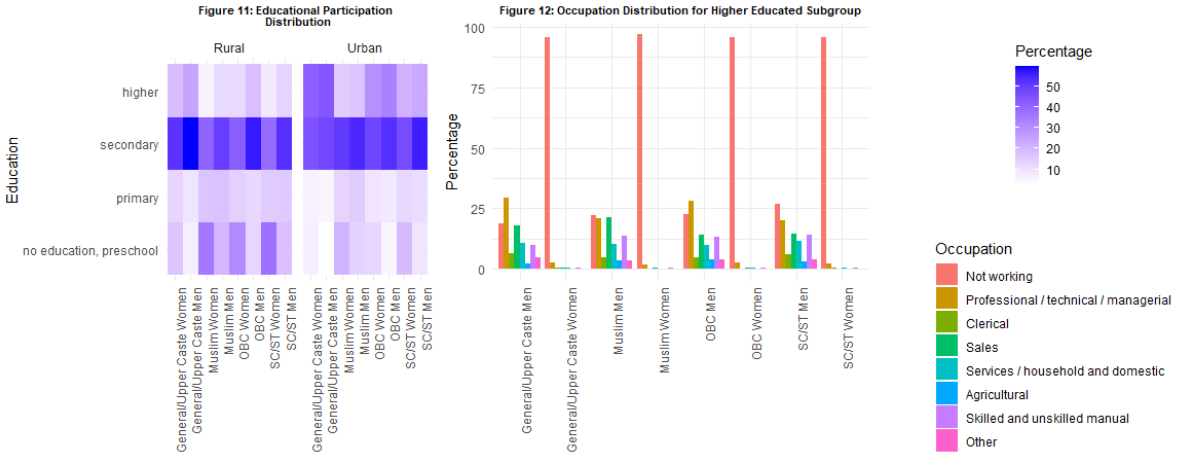
This indicates that inter-district inequality arises from distinctions in gender, caste, and religion. It is important to note that 'wealth' was not included as a covariate in this analysis. Despite this omission, most districts had an R-squared value above 0.6, and no other covariates were considered. Our objective was not predictive analysis but rather to understand the extent to which demographic differences contribute to inequality. We may not have evidence to confidently assert that wealth is a major factor influencing inequality in educational participation at the primary or secondary levels, not as much as caste or religion factors in India. In the following section, we will investigate wealth mobility, its relation to education and occupation, and how these dynamics differ between communities.

3.3 Wealth and Social Class Transitions: A Markov Chain Analysis

One fundamental question that many economists over the years have addressed is ' Does educational participation provide equal benefits across different genders and communities?' (Hanushek & Woess-

mann,2008; Thorat & Attewell, 2007; Mullainathan & Bertrand, 2004). If education does not confer equal advantages, do communities lack the incentive to pursue higher education? In this section, we will explore these questions. Importantly, we will examine how net wealth mobility differs between high and low educational performing clusters, as well as between homogeneous spatial clusters and districts that are heterogeneous among their neighbors.

3.3.1 Exploratory Results : Key Summary Statistics



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 function 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.

In Table 3 we see 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 Table 3, 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.

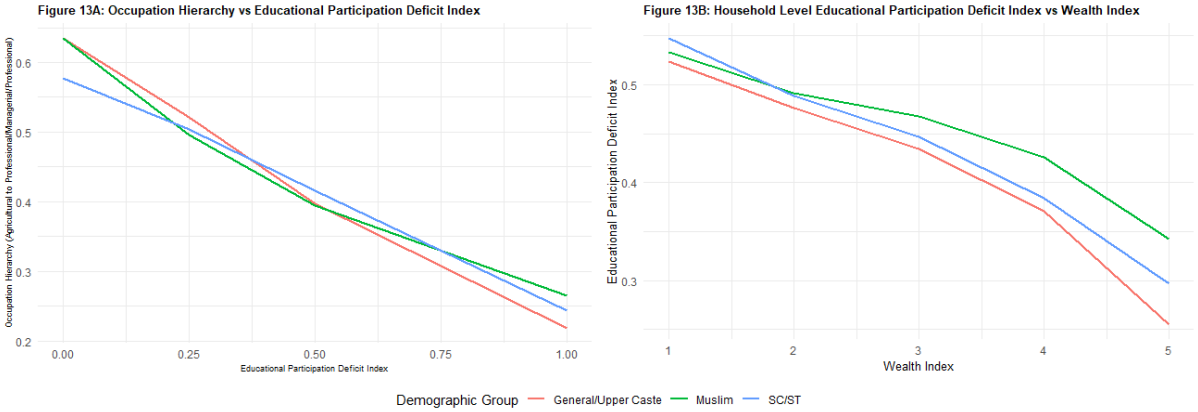
In summary, these socio-cultural determinants—caste, religion, geographic locality (rural or urban), and gender—persist in their influence, demonstrating that systemic structures and power dynamics are deeply embedded within the labor market. While it is evident that discrimination is more acute in rural 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 :

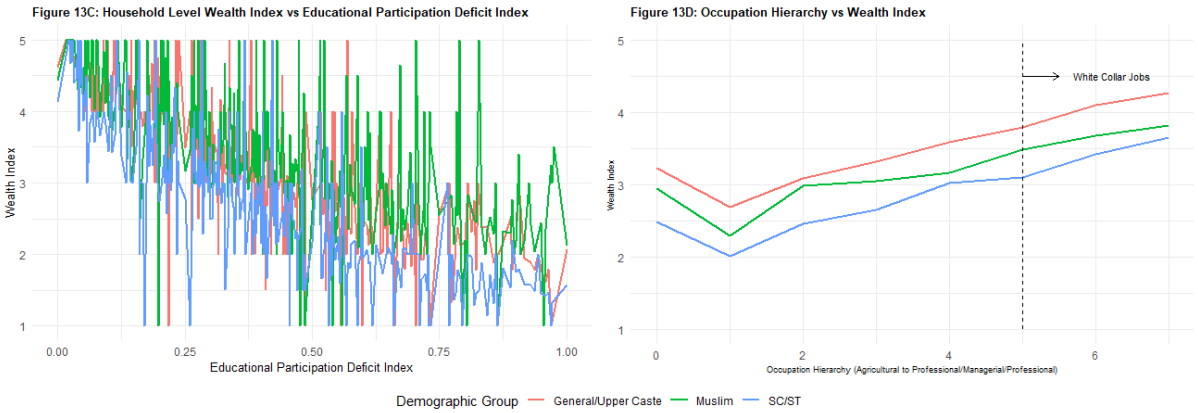
Table 3 Distribution (%) of Occupations by Demographic Group and Education Level

Urban Rural	Gender	Education Group	Demographic Group	Not Working	Professional/ Technical/ Managerial	Clerical	Sales	Services/ Household/ Domestic	Agricultural	Skilled/ Unskilled Manual	Other
Urban	Male	Higher	General/Upper Caste	20.5	28.8	5.4	16.12	10.21	3.06	11.36	4.32
			Muslim	21.9	20.8	4.8	21.25	10.16	3.32	13.86	3.69
			SC/ST	26.7	20.0	6.0	14.53	11.52	3.01	14.12	3.94
		Primary, Secondary	General/Upper Caste	7.8	5.1	2.4	19.59	12.01	9.33	37.78	5.83
			Muslim	7.6	3.2	1.4	19.13	10.30	6.09	43.77	8.33
			SC/ST	9.0	4.2	2.2	11.89	12.98	9.95	43.41	6.30
		No education, Preschool	General/Upper Caste	3.6	1.4	0	14.59	11.67	15.81	47.44	5.30
			Muslim	2.9	0.4	0.9	14.22	8.80	12.18	52.59	7.90
			SC/ST	5.10	1.07	1.61	6.72	12.903	15.86	50.80	5.91
	Female	Higher	General/Upper Caste	95.62	2.54	0.27	0.36	0.52	0.08	0.39	0.17
			Muslim	96.72	1.93	0.14	0.41	0.20	0.02	0.32	0.20
			SC/ST	95.67	2.38	0.31	0.22	0.53	0.13	0.45	0.27
		Primary, Secondary	General/Upper Caste	95.92742043	0.382138372	0.066807	0.662729	0.796344299	0.478341	1.429678523	0.25654
			Muslim	97.583421159	0.147523709	0.0281	0.351247	0.463645943	0.337197	0.927291886	0.161574
			SC/ST	95.383843345	0.288509784	0.094079	0.476668	1.204214752	0.696186767	1.473908868	0.382589

		No education, Preschool	General/Upper Caste	94.33934595	0.029329814	0.043995	0.542602	1.305176712	1.3491714	2.1850711 25	0.205309
			Muslim	97.02991453	0.021367521	0.021368	0.277778	0.662393162	0.5128205	1.2393162 39	0.235043
			SC/ST	93.29857088	0.077249903	0.07725	0.405562	1.487060641	1.9119351	2.2981846 27	0.444187
Rural	Male	Higher	General/Upper Caste	25.99831862	15.42664985	3.488861	8.51198	5.212274065	28.331232	9.4787725 94	3.551913
			Muslim	30	20.75757576	2.878788	11.36364	8.636363636	12.424242	10.606060 61	3.333333
			SC/ST	29.38508065	14.21370968	2.872984	5.897177	6.098790323	25.856855	12.600806 45	3.074597
		Primary, Secondary	General/Upper Caste	7.214342727	2.050715294	1.178854	7.331	5.470620744	49.54258	23.785841 47	3.426045
			Muslim	7.708665603	2.020202020	1.222754	11.80223	6.485911749	31.472621	33.599149 39	5.688464
			SC/ST	7.545330474	1.618249171	0.935855	4.893741	5.069214272	47.455644	28.436342 37	4.045623
		No education, Preschool	General/Upper Caste	3.643410853	0.116279070	0.116279	2.713178	3.178294574	65.271318	22.480620 16	2.48062
			Muslim	3.406113537	0.698689956	0	5.58952	4.541484716	45.502183	35.720524 02	4.541485
			SC/ST	2.850877193	0.255847953	0.219298	2.302632	2.850877193	63.413743	24.634502 92	3.472222
	Female	Higher	General/Upper Caste	96.37506857	1.675317551	0.118158	0.143478	0.379794911	0.74271	0.3966746 85	0.168798
			Muslim	96.88744119	1.375316685	0.072385	0.072385	0.398117988	0.2533478	0.8324285 2	0.108578
			SC/ST	95.62996595	1.362088536	0.181612	0.249716	0.488081725	1.3053348	0.5561861 52	0.227015
		Primary, Secondary	General/Upper Caste	95.34763402	0.169140765	0.055717	0.254706	0.360170335	2.8236558	0.8049110 52	0.184065
			Muslim	97.22694519	0.125495911	0.044531	0.178123	0.441259817	0.9715812	0.8217958 06	0.190268
			SC/ST	94.27330174	0.233532523	0.063535	0.231815	0.479085102	3.5459166	0.9100899 79	0.262724
		No education, Preschool	General/Upper Caste	93.82525538	0.034775049	0.021734	0.145621	0.245598783	4.8315584	0.7650510 76	0.130406
			Muslim	96.51093636	0.013216153	0.072689	0.178418	0.389876429	1.8766933	0.7599286 33	0.198242
			SC/ST	92.75304749	0.042205616	0.02731	0.158892	0.292956628	5.5587279	0.9508676 98	0.215993

We recall the measurement indication we created in section 2.3.1 and examine the relation between each of those pair of indicators (in Figure 13A-13D) .





3.3.1 Markov Chain Analysis

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.

The general expression we derived in section 2.3 was :

P_{w_j,w_i} =

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

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

For SC/ST, Muslim and General/Upper Caste Category the three Transition Matrices we get are :

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

	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

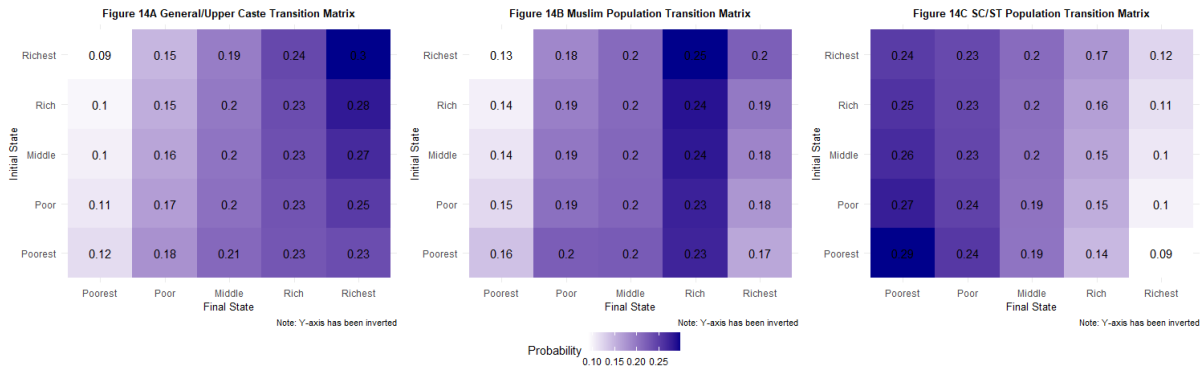
$P_{\text{muslim}} =$

	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

$P_{\text{general}} =$

	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

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



Note : 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%.

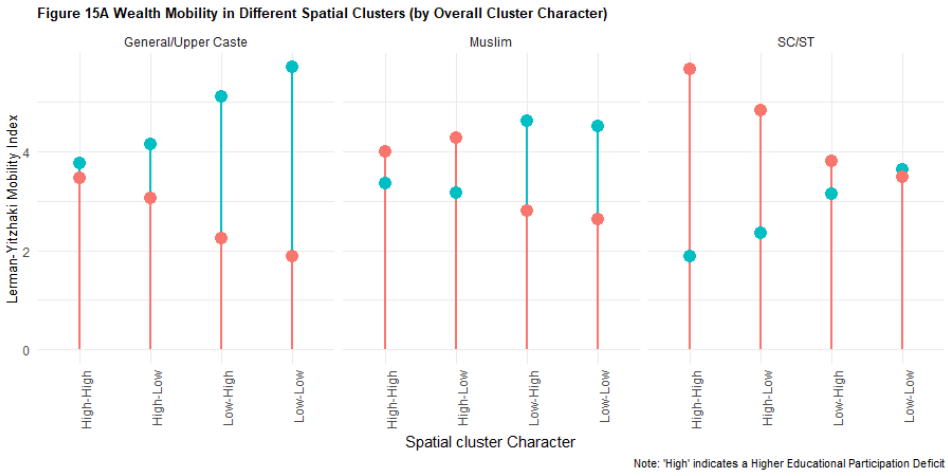
The SC/ST population state transition matrix highlights the challenges faced by this group. The transition probability of moving from the "Poorest" state to the "Poor" state is the highest among the three groups, reflecting significant barriers to upward mobility. The highest absorbing state probability 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.

3.4.1 Correlation Between Wealth Mobility and Educational Participation

There are established methodologies for evaluating the returns on education. Notably, the standard Mincer equation, articulated by Mincer (1974) and Becker (1964), examines the influence of additional schooling and work experience on wages. It scrutinizes the effect of an extra year of education and the incremental work experience on earnings. This equation includes a parameter for the rate of return on educational investments and incorporates work experience in a quadratic form to capture the diminishing returns often observed over time (Fleischhauer, 2007). However, the pursuit of educational qualifications carries profound economic implications, necessitating both temporal and financial investments. According to the most recent Periodic Labour Force Survey, returns to education in India are unequally distributed among different communities. Our inquiry delves into whether regional contexts exert an influence on this phenomenon. We explore whether communities make educational decisions—ranging from primary and secondary to higher education—based on the spatial clusters to which they belong, and how these decisions are shaped by their regional affiliations.

We commence our analysis by first (a) examining upward and downward mobility in relation to the inherent characteristics of the spatial clusters (the four types of clusters) (b) then investigating the variations within each spatial cluster as defined by their cluster ID (detailed in section 3.1.2), (c) next performing district-wise correlation mapping between educational participation and long-term wealth mobility, and (d) ultimately presenting maps that depict how the nature of educational participation aligns with the patterns of net social mobility. In Figure 15A to 15C in the axis we have the spatial clusters and in the Y axis the LY mobility Index. (defined in section 2.3).



The essential aspect to consider in Figure 15A (as well as figure 15b and 15C) is that the mathematics of these spatial clusters did not incorporate net mobility or clustering based on mobility metrics. Rather, the clustering emerged from a general assessment of overall educational participation. We probed whether zones of low educational participation correspond to diminished mobility and vice versa. Furthermore, we explored mobility within low-high regions—where areas with lower

participation deficits (higher participation) are bordered by higher deficit districts—versus low-low zones (spatial clusters of high participation) . The highest mobility was observed in clusters of higher participation, suggesting that clustering exerts more influence than being an isolated high or low-performing area. These spatial clusters were also not computed separately for different demographics, yet distinct patterns emerged within demographic communities, illuminating the broader relationship between mobility and educational attainment across communities

Figure 15B Wealth Mobility in Low-Low Spatial Clusters (High Educational Particpation Clusters)

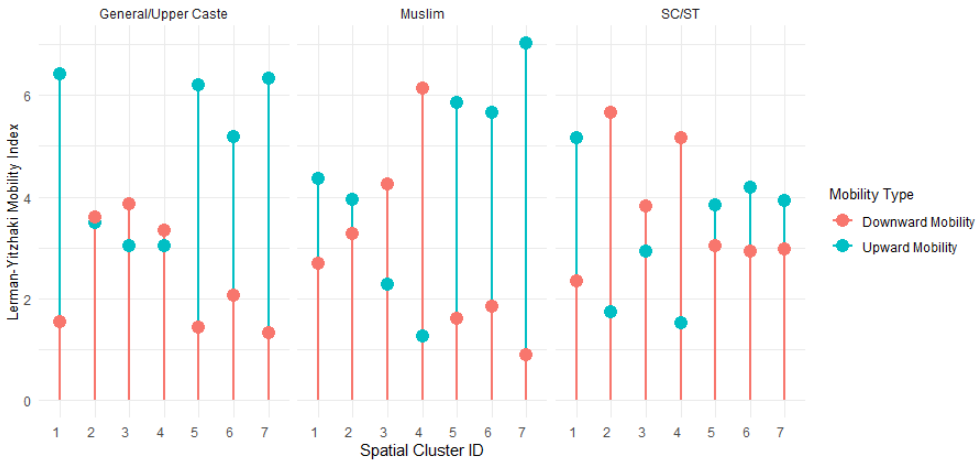
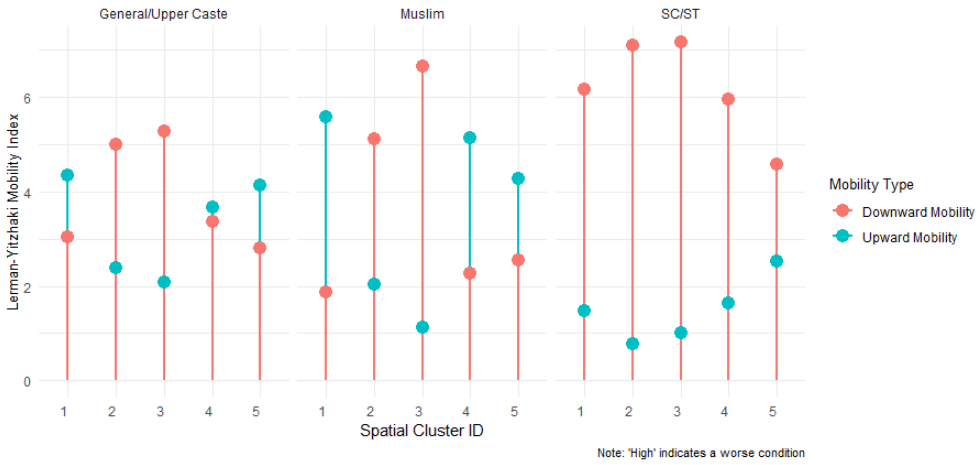


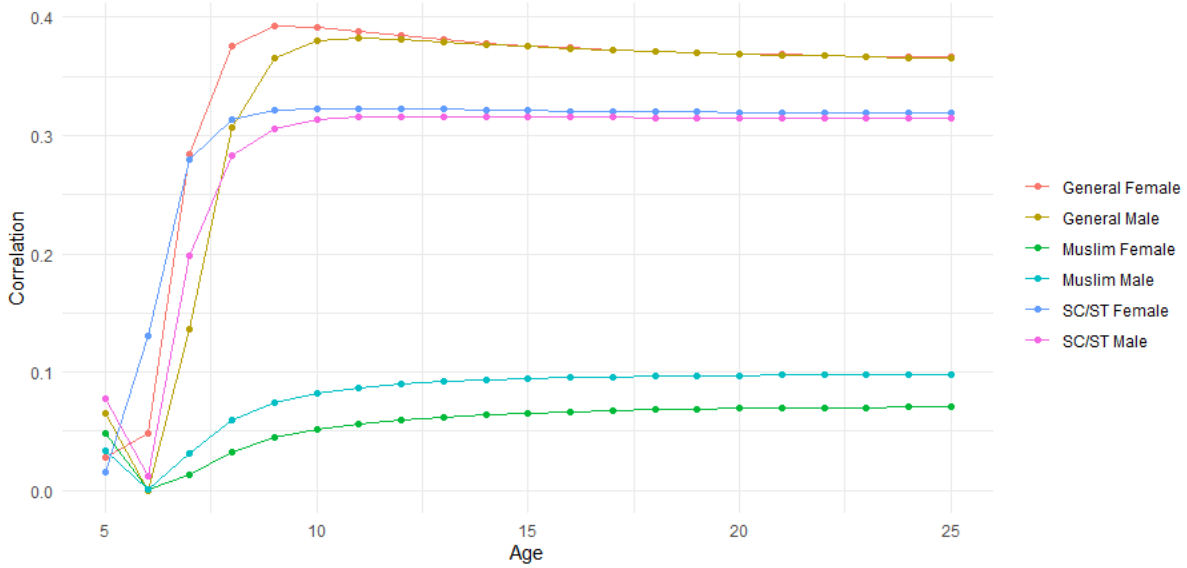
Figure 15C Wealth Mobility in High-High Spatial Clusters (Low Educational Particpation Clusters)



It is now evident that in nearly all regions with low educational participation, the SC/ST community experiences downward mobility, which may contribute to their lower participation rates. The broader regional trends in educational participation, whether high or low, significantly impact communities as a whole. Notably, the SC/ST community, which our Markov Chain analysis identified as having the highest steady-state probability of becoming the poorest under the current framework, shows signs of positive mobility within high educational participation clusters. This observation is particularly encouraging and suggests that increased educational participation can foster upward mobility for marginalized groups.

However, one notable observation is also that no clear pattern has emerged for the Muslim community. Specifically, as shown in Figure 15A, the Muslim community exhibits higher downward mobility in high-low districts compared to high-high spatial clusters of districts. In our forthcoming and final analysis of this section, prior to presenting the explicit map for the spatial clusters, we will examine the correlation between the number of years of education across different age groups and their correlation with net mobility within each district.

Figure 16 Correlation between Regional Net Mobility and Number of Years of Education



What emerges is that higher net mobility significantly influences the number of years of education. Wealth mobility demonstrates a stronger correlation with educational participation than the wealth index. However, as observed in Figures 15A to 15C, educational participation for the Muslim community does not appear to depend on mobility, nor does mobility depend on educational participation. This indicates that the Muslim community remains in a stagnant situation even when they attend school or college, leading to little incentive to pursue education since their occupational clustering likely does not involve white-collar jobs. Similarly, while the SC/ST community also experiences low net mobility, certain regions show reduced discrimination, which coincide with areas of high educational participation.

3.4.2 Education Participation Cluster Maps

We identified 12 spatial clusters, aiming to pinpoint districts that resemble their neighbors and share historical geographical realities shaping a common human experience. Utilizing the Educational Participation Deficit Indicator, we delineated these regions in section 3.4.1. Our analysis reveals that these clusters exhibit similarities not only in educational participation but also in inter-community social wealth mobility, especially among SC/ST and General/Upper Caste communities. While it is

evident that further examination of additional human development indicators is necessary to establish these clusters as more valid than conventional state borders, such an investigation is beyond the scope of our current study with the NFHS-5 data. In this section, we present the maps of three low participatory regions and four high participatory regions.

Figure 17A Educational Participation (High) Cluster ID -1

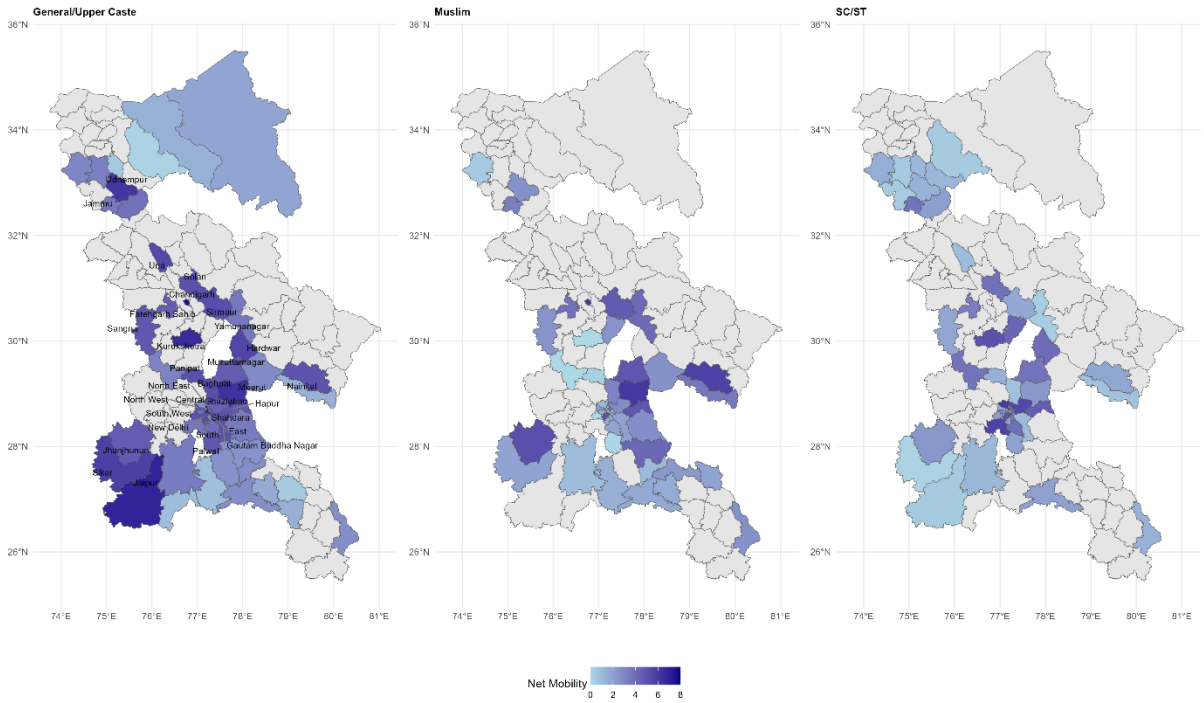


Figure 17B Educational Participation (High) Cluster ID -5

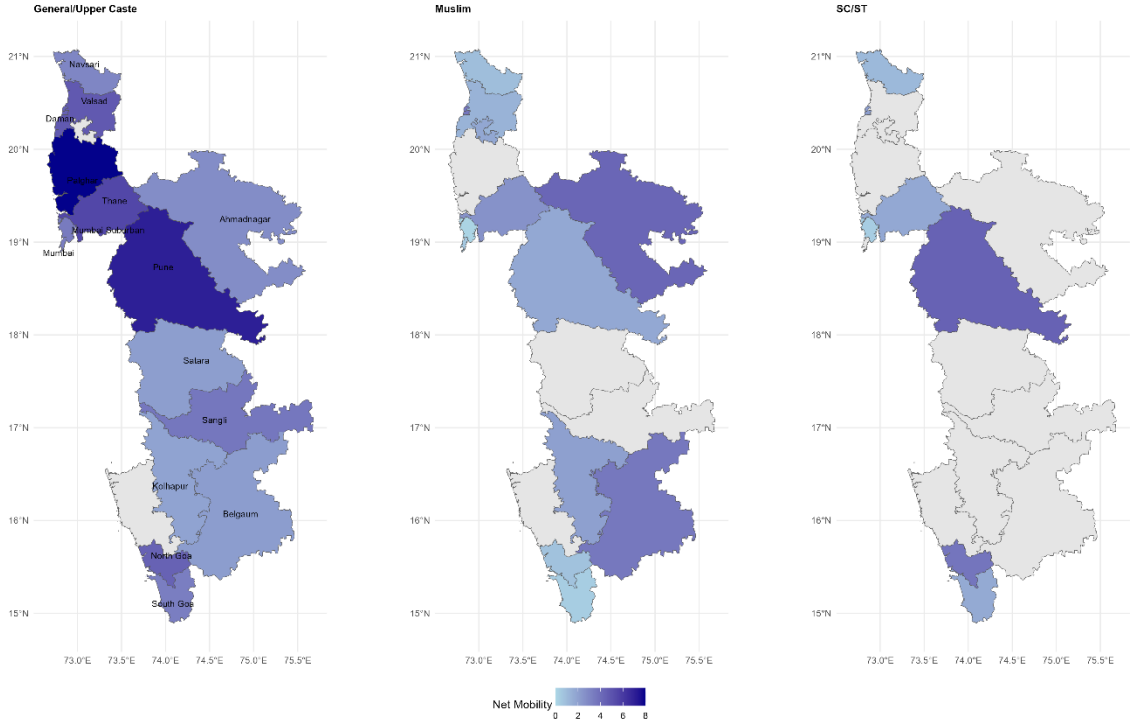


Figure 17C Educational Participation (High) Cluster ID -6

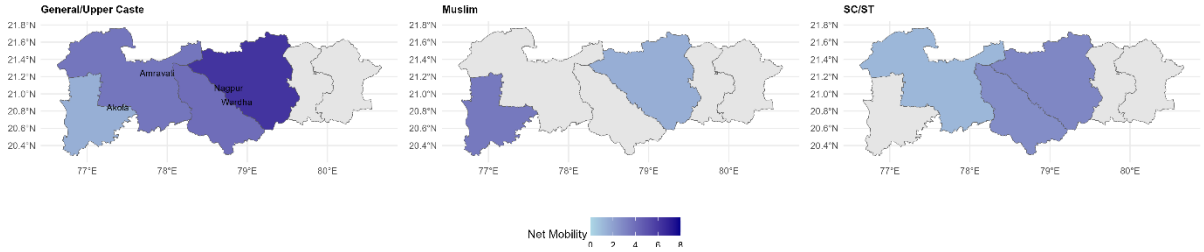


Figure 17D Educational Participation (High) Cluster ID -7

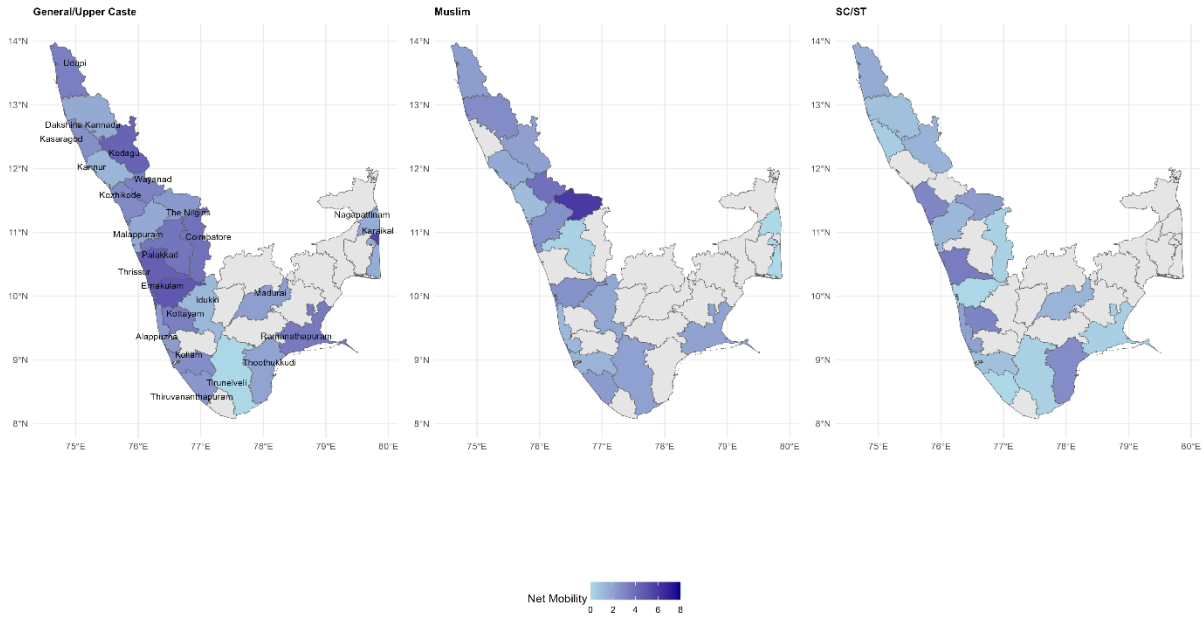


Figure 17E Educational Participation (Low) Cluster ID -1

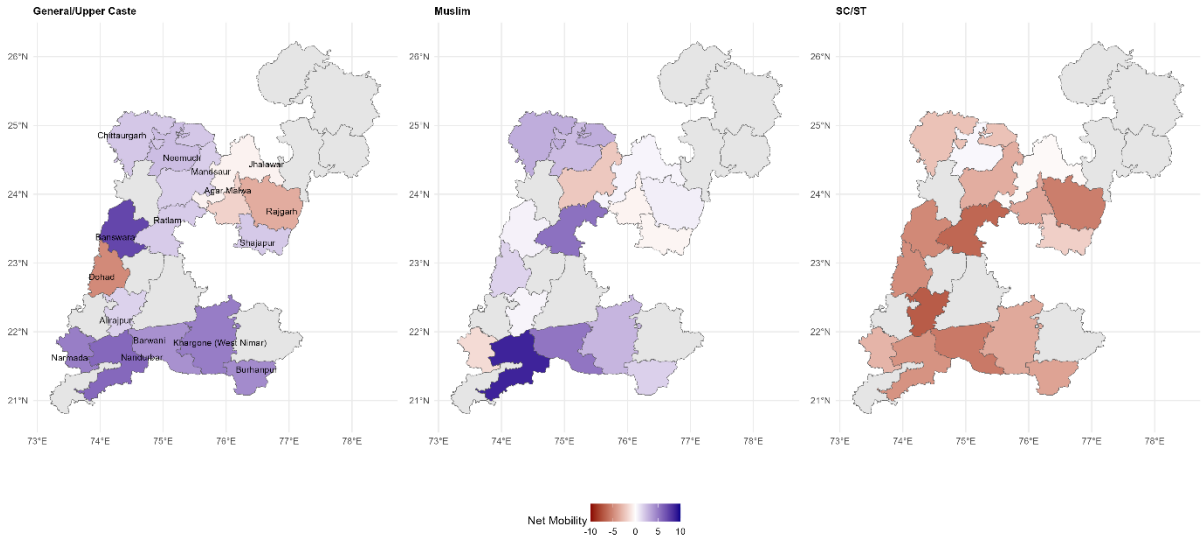


Figure 17F Educational Participation (Low) Cluster ID -2

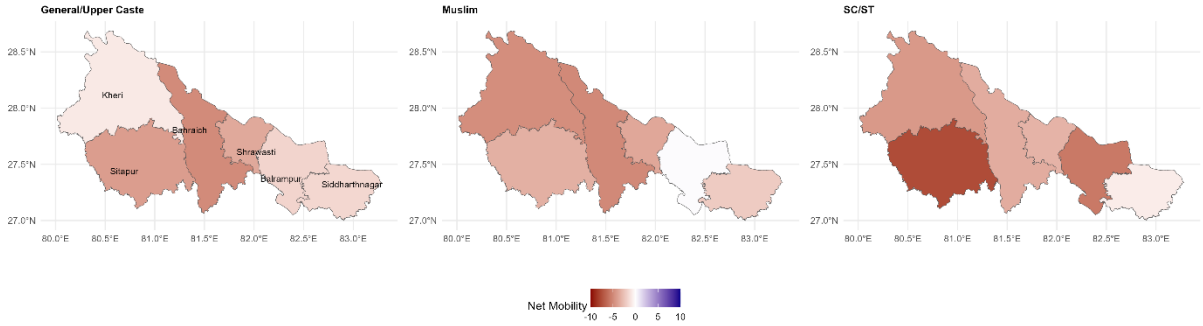
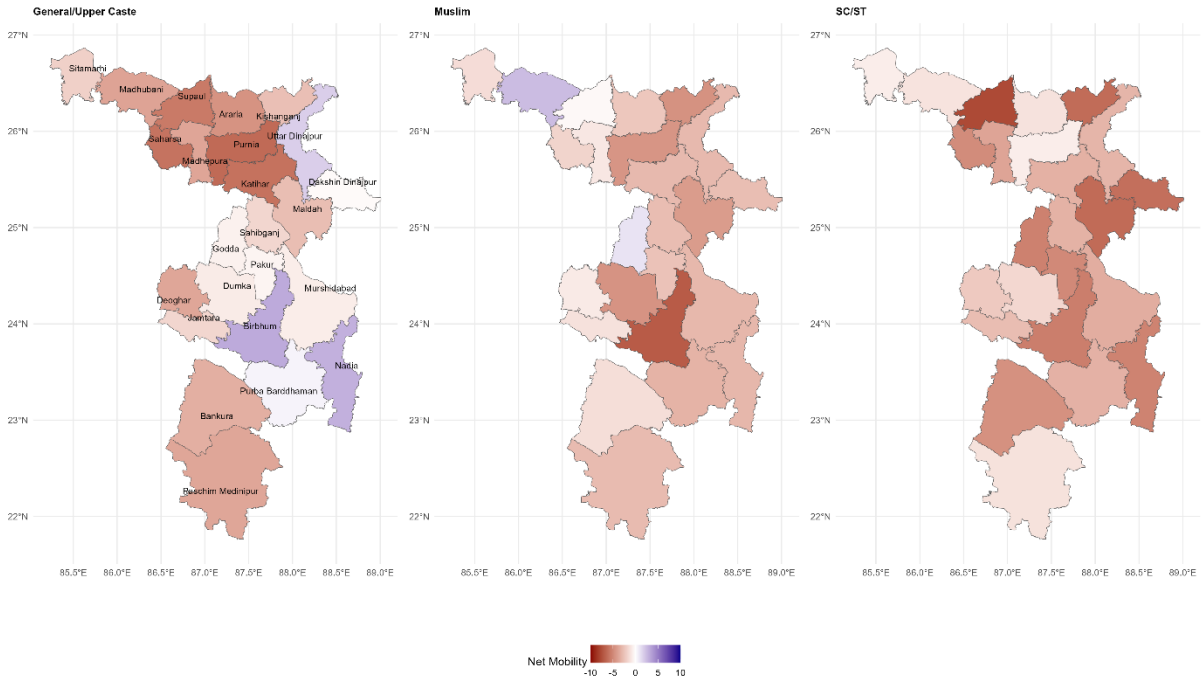


Figure 17G Educational Participation (Low) Cluster ID -3



3.5 Logistic Regressions : Comparing other Human Development Indicators Before and After Controlling for Education

The following section introduces the Empirical Models, which aim to estimate the relationship between an outcome variable (Y) and predictors (X), emphasizing the effect of education alongside demographic factors. The source of dependence may arise solely from demographic factors or may lose significance when education is controlled. One prevailing belief is that the pronounced effects of caste and religion result from certain demographic groups' inability to overcome historical disadvantages. Consequently, even with higher education, they may not be sufficiently skilled for the labor market. The wealth mobility disparities we observed are often blamed on the communities themselves. However, an alternative hypothesis suggests that it is not their lack of fitness upon receiving higher education, but rather society's employers' persistent assumption that these candidates are not meritorious or sufficiently educated. Here, we examine whether education

positively influences other human habits, civic sense, or financial behavior. This analysis will help us evaluate the validity of these competing hypotheses.

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)

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)	-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 EdI_inv	With EdI_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.