Community Effect on Educational Participation in Contemporary India: A District Level Analysis

Arnab Samanta

WORK IN PROGRSS PLEASE DO NOT DISTRIBUTE

Abstract

Existing studies on community, gender effects, and educational disparities in India have largely relied on state-level analyses of administrative data, often using fixed effects and universal community effect assumptions (common slopes) for caste-religion-related disparities. While these approaches offer valuable insights, they obscure critical intra-state heterogeneity, introduce bias, and reduce statistical power to detect local variation. This study addresses these challenges by incorporating district-level regressions and Bayesian posterior adjustments, capturing localized community effects that state-level analyses fail to identify. The findings are supported by prior village-level RCTs and suggest that wealth determinants, often emphasized as superseding community effects, may themselves be endogenous. Markov steady-state wealth transition probabilities further reveal caste-based disparities, underscoring the need for deeper exploration. By integrating spatial dimensions and advanced methodologies, this study provides a comprehensive framework for understanding educational inequalities in contemporary India.

1. Introduction

The literature on the economics of education has consistently emphasized parental education as an important indicator of children's educational attainment (Shavit and Blossfeld, 1993; Rosenzweig and Wolpin, 1994; Behrman and Rosenzweig, 2002; Treiman et al., 2003; Hertz et al., 2007). Given that foundation, empirical patterns of disparities in educational enrolment across caste and religious groups in Indian society are not unexpected. Seminal studies such as, Desai and Kulkarni (2008), using National Sample Survey (NSS) data from 1983, 1987–1988, 1993–1994, and 1999–2000, report that by 1999–2000, 37% of Dalits, 44% of Adivasis, and 32% of Muslims among males aged 24–29 had never attended formal schooling, compared to only 17% of upper-caste Hindus and other religious groups. More recent studies that analyzed National Family Health Survey (NFHS-5) data states : maternal literacy rates are 83% among forward-caste Hindu mothers compared to 51% for SC-ST mothers, who average only 5.26 years of schooling versus 9.47 years for their forward-caste counterparts (Deshpande and Ramachandran, 2021). However, the community effect, while observable through the lens of parent and child education link, can be shaped by a complex interaction of social identity, economic challenges, and policy measures. Several studies leveraging IDHS, NSS, and NFHS data have concluded that these inter-community disparities are largely attributable to differences in wealth or other exogenous factors, which they often do not conceptualize as pre-determined endowments (Jalan and Murgai, 2008; Maitra and Sharma, 2009; Emran and Shilpi, 2012; Hnatkovskay et al., 2013). On the other hand, contemporary economics literature emphasizes that a person's school or college enrolment decision is deeply embedded within broader structural contexts —social networks, regional access and infrastructure, and political-legal frameworks (Novosad 2022; Bailwal and Paul 2019,2020, 2021). Consequently, relying on any single factor to assert the absence of discrimination proves inadequate.

That caste and religion are an important factor explaining inequalities not only in education but in the labor market, earning gap, global exposure has been attenuated by influential studies (Thorat and Attewell 2007 , Vakulabharanam, 2010, 2011). Specifically focusing on educational enrolment while there are works that used NSS, NFHS, IDHS or other administrative data, on the other hand we have influential research works since the early 2000's that conducted their novel fieldworks. (Jefferey et al. Iyer Borooah and Iyer , Hoff and Pandey) .

This study seeks to evaluate the community effect on educational attainment in contemporary India using the latest survey data (NFHS-5 2019-21) . However, to address this seemingly straightforward question which has already been extensively researched (with older rounds of NFHS and other sample survey and census data), we propose methodological improvements. First, for the outcome variable, we introduce an age-adjusted educational attainment scale (0-1) that integrates years of education and levels of schooling. Second, we demonstrate that community effects vary spatially within the country. Pooling data across states for a single regression masks these variations . Hence the use of state dummies with common slopes in national-level regressions that assumes homogeneity within states, reduces statistical power. and. Third, India's demographic diversity underscores the necessity of district-level analyses.

For example, districts such as Bangalore (9.63 million), North 24 Parganas (10.08 million), and Pune (9.43 million) have populations exceeding that of Switzerland (8.96 million). Between districts regional variations driven by residential segregation (Novosad 2022), recourse allocation (Jhingran & Sankar 2009) and local cultural norms, make district-level modelling a minimal analytical necessity. By conducting 707 district-specific regressions, we reveal significant variability in the magnitude and presence (or absence) of community effects, with these varying slopes across districts best visualized through maps. Fourth, this leads to another estimation challenge: while district-level analysis captures spatial heterogeneity, districts are not isolated entities. Spatial dependencies exist as well. To address this, we employ spatial autocorrelation tools and Empirical Bayesian posterior estimates, framing the analysis as a compound decision problem. This approach identifies spatial clusters that traverse state boundaries, revealing that neighbouring districts from different states may exhibit similar educational patterns, while differing significantly from other districts within their own states. This novel approach also provides an alternative to Geographically Weighted Regression (GWR), offering a more practical solution for complex studies. Fifth, we propose accounting for household-level random effect which is rarely considered in studies that have previously used household-level microdata such NFHS or NSS. Mixed-effects models can account for the potential correlations between observations drawn from the same household and mitigating heteroscedasticity concerns. Finally, to resolve the Wealth vs. Caste/Class Debate, we examine wealth transition probabilities across caste and religious groups (e.g., forward caste, SC/ST, Muslims). By calculating Markov steady-state probabilities, we enable meaningful comparisons of transition matrices and demonstrate that wealth mobility itself is intrinsically tied to caste, emphasizing that wealth and caste effects are not orthogonal.

These methodological advancements collectively yield a more robust model.

2. Methodology

In this section we will provide methods for constructing I) Age-adjusted Educational Attainment (EA) II) Spatial Clusters III) Inter-cohort Educational Mobility : Gini coefficient induced III) A mixed-effect district level multistage regression model to compare determinants of educational participation but most importantly to see the role wealth IV) Quartile Earnings Transition Matrices for General/Forward, Muslim, SC/ST communities and respective Markov Steady State Probabilities.

I) Age-adjusted Educational Attainment (EA). In the NFHS 2019-21 data we get multiples 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. For our analysis, we computed an age-adjusted educational participation indicator, building on criteria similar to those used in the Multidimensional Poverty Index (MPI) framework developed by Alkire and Foster. These criteria, have also been adopted by the Indian government for its MPI calculations. The specifics of this calculation are outlined in Table 1.1.

In the ideal scenario, Educational Attainment (EA) is assigned a value of 1. For example, where a child under 12 is enrolled in school, an adolescent aged 12-18 is attending school and has started/completed secondary education with at least six years of schooling, or an adult has partially or fully completed college education. In contrast, EA is assigned a value of 0 for instances of complete educational deprivation (such as when a child is not in school, or an adolescent/adult has fewer than six years of schooling.)

Table 1 : Age-Adjusted Educational Attainment(EA) Measurement

Age Group	Combined Educational Participation Observation from NFHS 2019-21	Raw poverty Score (aligned with MDG criteria)	Educational Attainment (EA)
Children (5-12 years)	Not currently attending school	4	0
	Currently attending school or has attended school during the survey year	0	1
Adolescents (12-18 years)	Not attending school and has less than six years of schooling	4	0
	Not attending school but has six or more years of schooling, though has not completed secondary education	3	0.25

Individuals (18 years and above)	Attending school, has six or more years of schooling, but still pursuing primary education	2	0.5
	Not attending school but has completed secondary education	1	0.75
	Attending school, has started or completed secondary education but not higher education, and has six or more years of schooling	0	1
	Less than six years of schooling	4	0
	Six or more years of schooling, completed primary education but has not pursued secondary education, and is currently not out of school	3	0.25
	Six or more years of schooling, started secondary education but has not completed it, and is currently not attending school	2	0.5
	Six or more years of schooling, completed secondary education but not higher education, and is currently not attending school	1	0.75
	Attending school and has completed secondary education or higher, or not attending school but has pursued higher education	0	1

Since the variables we analyze to construct the age-adjusted metric, are measured using the same methods of schooling level and data- so we would not have any compatibility issues.

II) Spatial Clusters. 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 X_i and X_{J_i} such that $X_{J_i}$are the values observed in the neighborhood J_i of the i^{th} spatial entity.

Anselin proposes a LISA measure as, $LISA_i = \frac{x_i - \bar{x}}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \sum_{j \neq i} w_{ij}\left(x_j - \bar{x}\right)$

where, w_{ij} (spatial weight modelled over distance) $= K\left(d_{ij}, h\right) = exp\left(-\frac{\text{distance} \equiv d_{ij}}{\text{threshold} \equiv h}\right)$

the function $K\left(d_{ij}, h\right)$ is the Kernel function where h is the bandwidth parameter; we have taken 0.2. A small threshold value captures the local interactions better . The choice of exponential/Gaussian kernel is based on the hypothesis of gradual decay. Exponential Kernel further emphasizes faster decay. The magnitude of LISA_I indicates how strongly, positively or negatively the neighboring x values (in our case educational attainment) are correlated to the i^{th} spatial entity (in our case the i^{th} district) As , distance increases w_{ij}(spatial weight) diminishes to zero and thus very distant districts contribute nothing to the aggregated autocorrelation measure.

Figure 1

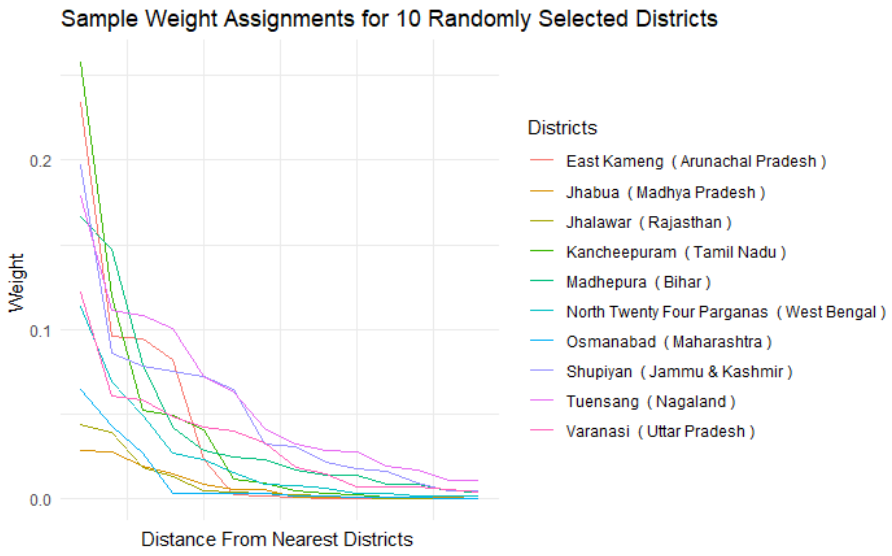


Figure 1 illustrates this rate of decay i.e. what a threshold of 0.2 means, and how the influence diminishes with increasing distance. For example, weights assigned to Jhalwar (Rajasthan), where neighboring districts are larger, sharply decay to zero approximately after the 5th district. In contrast, 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

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'. We can observe multiple clusters –while several districts may remain outside any cluster. In this study, each spatial cluster will be assigned a unique ID, accompanied by corresponding maps. For example, the 12 Indian districts of 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), and Wanaparthy (Telangana) form a spatial cluster. So, a spatial cluster may span across states (in the above example, Karnataka, Andhra Pradesh, Telangana).

III) Educational Attainment Inequality for two birth-cohorts. We calculate within-district Gini-coefficients in education participation for two birth cohorts 2001-2009 and 1950-1981 and we test for difference using influence function-based variance approximation for the Gini coefficient.

$$\widehat{Var}(\hat{G}) \approx \sum_{i=1}^{n} \left(\frac{(2i - n - 1)}{n} \cdot \frac{X_{(i)}}{\bar{X}} \right)^2$$

Where $X'_{(i)}s$ are the order statistics, and $\bar{X}$ is the sample mean.

Thus, if we have Two Gini Coefficients G1 and G2 we can test then with Z statistics

$$Z = \frac{\widehat{G_1} - \widehat{G_2}}{\sqrt{\widehat{Var}(\widehat{G_1}) - \widehat{Var}(\widehat{G_2})}}$$

IV) District-Level Regression Results with Bayesian Shrinkage.

Before we introduce the model, let us provide the background what happens when we overlook the regional disparity and consider fixed effect state dummy and in which way that may affect our result.

For simplicity let's say our true model for the d^{th} district is (discarding the household level random effect and between district borrowing possibility)

$$y_{i,d} = \beta_{k,d} X + \epsilon_{i,d} \quad , d = 1, \dots, D \qquad (1)$$

Where Each district d has its own slope. y and X are our independent and dependent variables.

Now if we instead take fixed state-dummy model which imposes a single slope across all districts, including state-level fixed effects our model would be :

$$y_{i,d} = \beta^{A}{}_{k,d} X + \delta_{s(d)} + \vartheta_{i,d} \quad , d = 1, \dots, D \qquad (2)$$

$\beta^{A}{}_{k,d}$ = aggregate national slope.

$\delta_{s(d)}$ = fixed effect for state *s(d)*

So, the model (2) actually assumes that $\beta_{k,d} X = \beta^{A}{}_{k,d}$ for all districts.

a) Aggregation Bias

$$E\left[\hat{\beta}_k^A\right] = \frac{\sum_{\{d=1\}}^{D} w_d \beta_{k,d}}{\sum_{\{d=1\}}^{D} w_d}$$

Where w_d are weights based on the distribution of X and y - the entire set of micro-level variables that enter into the aggregation process in the districts and that is not in general equal to $\frac{1}{D}\sum_{\{d=1\}}^{D} \beta_d$. (Theil, 1954)

b) Reduced Power

The residual term $\vartheta_{i,d}$ now includes both $\epsilon_{i,d}$ and a misspecification component $\beta_{k,d}X - \beta^{A}{}_{k,d}X$. (Since the state fixed effects are chosen to best fit the intercept shifts (i.e., they are estimated to remove systematic level differences across states)

$$SE\left(\hat{\beta}_k^A\right) \propto \sqrt{\frac{\sum_{\{d,i\}} \vartheta_{i,d}{}^2}{\sum_{\{d,i\}}(X_{\{i,d\}} - \bar{X})^2}}$$

Since ignoring heterogeneity inflates the denominator in the t-statistic for testing H0: $\beta^{A}{}_{k,d} = 0$, now involves a larger residual variance. Even if $\beta^{A}{}_{k,d}$ is close to some average of the true slopes, the heightened variance makes it harder to reject the null hypothesis, thus reducing statistical power.

c) Household Level Random Effect.

On the other hand, the household level correlated observations would increase the chance of Type-I error which is the opposite problem of what we discussed in the previous point.

So, the model we propose is the following. It has two stages – Stage I) 707 district level mixed-effect regressions. Stage II) We recall the clusters we created in section 2(b), and we shrink the parameter using Empirical Bayesian shrinkage within these clusters. We follow the James-Stein suggestion

Stage – I

Because of the above challenges we will run 707 separate regressions for each district. And we also

$$\textit{Number of Years Education}_{ijd}$$

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

Here, $\textit{demographic factor}_k \in \{Muslim, SC/ST, Female\}$ (Forward caste males as the baseline)

$k = 1,2,3$; $p = 4,5$

$\textit{Number of Years Education}_{ijd}$ ≡ for i^{th} individual from j^{th} household at the d^{th} district

$exp\left(\frac{-1}{age_{ijd}^2}\right)$ ≡ age- adjustment factor

$X_{ij,d} = \{\text{wealth quintile, urban /rural, other household characteristics}\}$ ≡ control variables

$u_{j,d} \sim N(0, \sigma_u^2)$, Household level random effect for j^{th} Household at the dth district.
$\epsilon_{ij,d} \sim N(0, \sigma^2)$, $k \in \{Muslim, SC/ST, Female\}$, $p \in \{Muslim * female, SC/ST * female\}$

The above equation has two other variations depending on where adequate data on a specific demographic group is not available as discussed in section 2.

Stage – II

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

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

The best way to represent the results of these 707 separate regressions is by providing the estimates on a map.

IV) Quartile Earnings Transition Matrices.

To address the question, whether wealth and wealth mobility also could be different between communities compute the transition probabilities i.e. probability of starting in wealth quintile category W_i and moving to wealth quintile category W_j. This question can be decomposed into a) What is the probability of an individual from Wealth Group W_i achieving Education Level EA_i b) What's the probability of an individual with Education Level EA_i obtaining Occupation O_i? c) What's the probability of an individual in Occupation O_i accumulating Final Wealth W_j^f? Here occupations are classified into categories: O_1 = White Collar (Professional, Clerical, Sales), O_2 = Blue Collar (Household or domestic Services, Skilled/Unskilled Manual), O_3 = Agricultural, and O_4 = Not Working.

The calculation for each demographic community is as follows leveraging the total probability theorem.

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

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

Steady State Probabilities.

Since we have the transition matrices we can calculate Markov steady-state distribution to have comparison of the transition matrices. In the context of our transition matrices, that will indicate the long-term probabilities of individuals being in each wealth category unless there is any targeted intervention. 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.

3. Data

The National Family Health Survey (NFHS), conducted every five years, has been a foundational dataset for research in India – even for disciplines other than social sciences, despite critiques of earlier versions. It is administered by the Ministry of Health and Family Welfare (MoHFW) and the National Sample Survey Office (NSSO) and one of the largest sample surveys in the country. We use the latest NFHS-5 (2019-21) data for our study. Covering all states, union territories (UTs), and 707 districts, NFHS-5 collected data from 636,699 households, 724,115 women, and 101,839 men, offering comprehensive insights into population, education, employment, and geography such as estimates for indicators like the wealth index, preschool education, incomplete education, access to schooling, reasons for dropouts, and occupation history.

i. Stratified Two-Stage Sampling. Districts were stratified into urban and rural areas. Rural strata were sub-stratified by village population and the percentage of scheduled castes and tribes (SC/ST).

ii. Primary Sampling Units (PSUs). Villages and Census Enumeration Blocks (CEBs) were PSUs, sorted by women's literacy rates and SC/ST population percentages. PSUs were selected using probability proportional to size (PPS).

iii. Household Sampling. Each PSU contained 100-150 households, with 22 systematically selected per cluster.

iv. Data Collection. A total of 30,456 PSUs were selected, and fieldwork was completed in 30,198 PSUs. Surveys were conducted using four questionnaires—Household, Woman, Man, and Biomarker—translated into 18 languages and administered via Computer-Assisted Personal Interviewing (CAPI).

v. Training and Quality Control. Field coordinators were trained through Training of Trainers (ToT) workshops. Data collection was monitored by district coordinators, IIPS project officers, and senior staff. Quality control included daily data transfers, checks, and real-time feedback. NFHS-5 achieved response rates of 98% for households, 97% for women, and 92% for men.

Table 2 offers an overview of our constructed EA metric, along with sample sizes for each subgroup are presented, along with t-tests and F-tests to examine differences across religion and caste groups.

Table 2 : Sample Size and Summary Statistics

Urban/Rural	Age Group	Demographic Parameter	Demography	Sample Size	Mean EA (SD)	t Test Statistic	F Test Statistic
Urban	Below 12 years	Gender	female	56857	0.9744 (0.1580)	-0.77	

			male	62406	0.9751 (0.1559)		
		Caste	General/Others	31031	0.9810 (0.1365)		34.896***
			Other Backward Class	50416	0.9733 (0.1611)		
			Scheduled Caste/Tribe	37816	0.9715 (0.1665)		
		Religion	Hindu/Jain	83874	0.9780 (0.1468)		127.921***
			Muslim	23017	0.9572 (0.2025)		
			Sikh	2179	0.9867 (0.1146)		
			Others	10193	0.9853 (0.1204)		
	12 to 18 years old	Gender	female	37903	0.8778 (0.2882)	6.88***	
			male	41300	0.8633 (0.3035)		
		Caste	General/Others	21090	0.9085 (0.2536)		267.047***
			Other Backward Class	33127	0.8640 (0.3024)		
			Scheduled Caste/Tribe	24986	0.8462 (0.3179)		
		Religion	Hindu/Jain	55995	0.8890 (0.2765)		610.109***
			Muslim	14817	0.7782 (0.3673)		
			Sikh	1581	0.8987 (0.2653)		
			Others	6810	0.9098 (0.2453)		
	18+ years old	Gender	female	227164	0.4698 (0.3892)	-99.359***	
			male	221589	0.5809 (0.3588)		
		Caste	General/Others	134729	0.6184 (0.3663)		6386.097***
			Other Backward Class	184729	0.4975 (0.3772)		
			Scheduled Caste/Tribe	129295	0.4650 (0.3751)		
		Religion	Hindu/Jain	335547	0.5409 (0.3791)		2466.669***
			Muslim	66356	0.4107 (0.3686)		
			Sikh	10557	0.5508 (0.3751)		
			Others	36293	0.5734 (0.3547)		
Rural	Below 12 years	Gender	female	219433	0.9621 (0.1909)	-3.395***	
			male	235255	0.9640 (0.1864)		
		Caste	General/Others	70425	0.9756 (0.1542)		332.345***
			Other Backward Class	174149	0.9665 (0.1799)		
			Scheduled Caste/Tribe	210114	0.9560 (0.2051)		
		Religion	Hindu/Jain	348120	0.9656 (0.1821)		397.699***
			Muslim	48262	0.9364 (0.2440)		
			Sikh	8753	0.9872 (0.1124)		
			Others	49553	0.9667 (0.1794)		
	12 to 18 years old	Gender	female	136971	0.7907 (0.3558)	-23.173***	
			male	139862	0.8211 (0.3343)		
		Caste	General/Others	44846	0.8712 (0.2917)		1370.437***
			Other Backward Class	108640	0.8152 (0.3376)		
			Scheduled Caste/Tribe	123347	0.7743 (0.3658)		

	Religion	Hindu/Jain	215825	0.8137 (0.3399)		727.232***
		Muslim	28837	0.7208 (0.3963)		
		Sikh	5852	0.8786 (0.2793)		
		Others	26319	0.8202 (0.3292)		
18+ years old	Gender	female	684389	0.2687 (0.3447)	-226.917***	
		male	638766	0.4073 (0.3578)		
	Caste	General/Others	249970	0.4278 (0.3697)		12371.23***
		Other Backward Class	509592	0.3362 (0.3589)		
		Scheduled Caste/Tribe	563593	0.2938 (0.3435)		
	Religion	Hindu/Jain	1044353	0.3354 (0.3593)		478.747***
		Muslim	109712	0.3058 (0.3542)		
		Sikh	35643	0.3760 (0.3554)		
		Others	133447	0.3501 (0.3478)		

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

For urban cohorts below 12 years, the tests indicate no significant differences in educational participation between males and females. Average educational attainment (EA) declines with age, particularly in rural areas, where the levels are notably lower. Generally, statistically significant differences are observed across categories, with rural ST adults and females emerging as the most deprived groups. Thus, in general the NFHS-5 large sample size ensures robust statistical power, allowing even minor deviations to achieve significance.

4. Results

The findings are presented in four main sections:

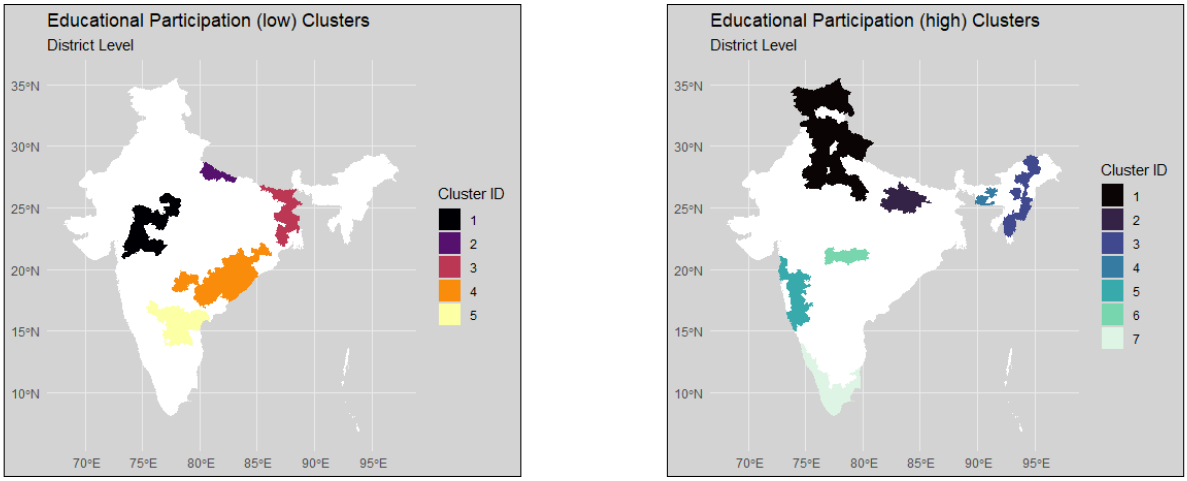
4.1) Between-Districts Spatial Autocorrelation: Spatial Clusters of Educational Attainment. In this section, we identify distinct geographic contours highlighting areas of both strong and weak educational performance. *4.2) Within-District Inequality Across Two Birth Cohorts.* We show evidence of a notable reduction in educational attainment (EA) inequality over time. To offer broader perspective, we also compare our results against cross-country Gini coefficients, illustrating that despite the improvements, inequality remains substantially high. *4.3) District-Level Regression with Bayesian Shrinkage.* By employing a regression model with varying slopes and a Bayesian shrinkage approach, we examine the influence of caste, class, and religion on EA. Our results indicate that these effects differ considerably across regions and are not uniform even within the same state. *4.4) Quartile Wealth Transition Matrix and Steady-State Probabilities.* Drawing on the 2019–21 NFHS-5 data, we analyse the wealth transition matrix and steady-state

probabilities for three groups—Forward Caste/OBC Hindus and other non-Muslim communities, Scheduled Castes and Scheduled Tribes (SC/ST), and those with Muslim religious backgrounds. This allows us to understand wealth mobility and persistence among these distinct communities.

(4.1) Between-Districts Spatial Autocorrelation: Spatial Clusters of Educational Attainment

In this section, our goal is to identify groups of districts that share similar levels of educational engagement, which is critical for the subsequent analysis. A total of 12 spatial clusters were identified. Interestingly, no cluster is composed solely of districts from a single state, which challenges the traditional state-level approaches on regional development dynamics. Additionally, some districts may not fall into any of the 12 clusters. For instance, cities like Kolkata, surrounded by regions with lower performance, prompt us to question the urban-centric approach to development.

Figure 2



This shed light on the regional reality that fluctuates between two primary aspects: the influence of socioeconomic, demographic, 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. Investments in one area can create spillover effects or even competition between neighboring regions.

(4.2) Within-District Inequality Across Two Birth Cohorts

Using the two sets of Gini coefficients for the birth cohorts 2001–2009 and 1950–1981, we apply asymptotic test statistics to evaluate differences. Figure 2 clearly demonstrates a decline in inequality from the older birth cohort. Additionally, testing the Gini coefficients yields p-values below 0.001 across all districts. Notably, this decline signifies not only an improvement in overall educational attainment but also a substantial reduction in inequality, signalling significant progress in enrolling children in school, particularly from groups that historically contributed to higher

Figure 3

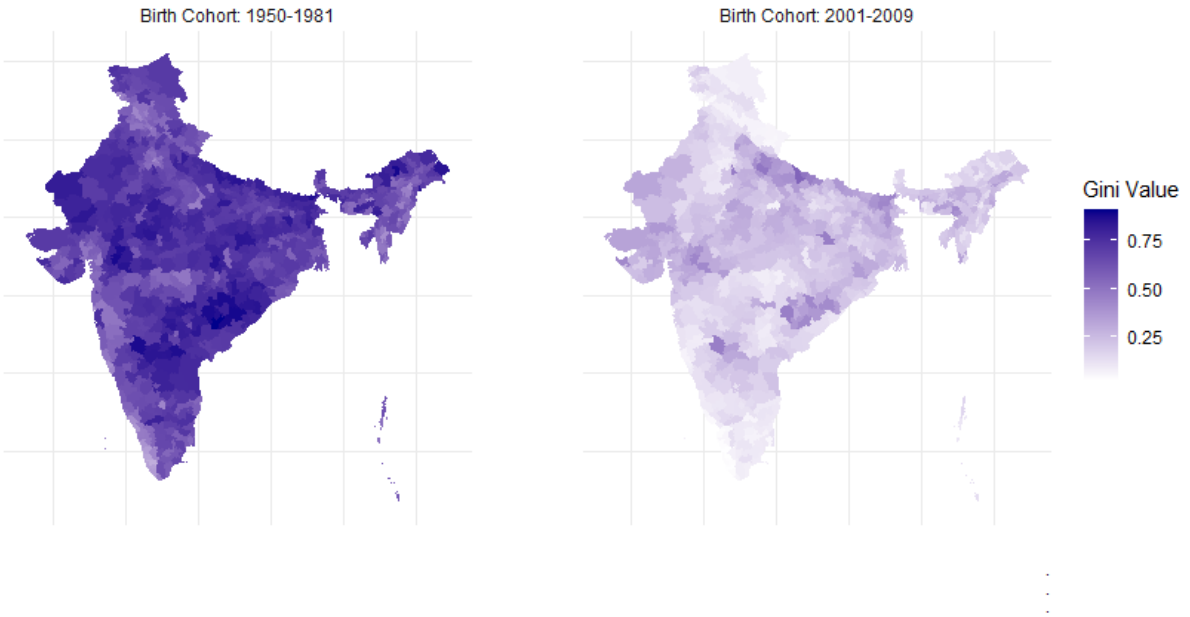
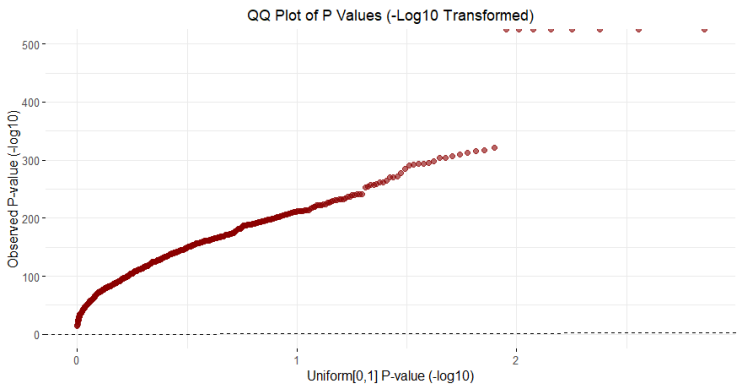
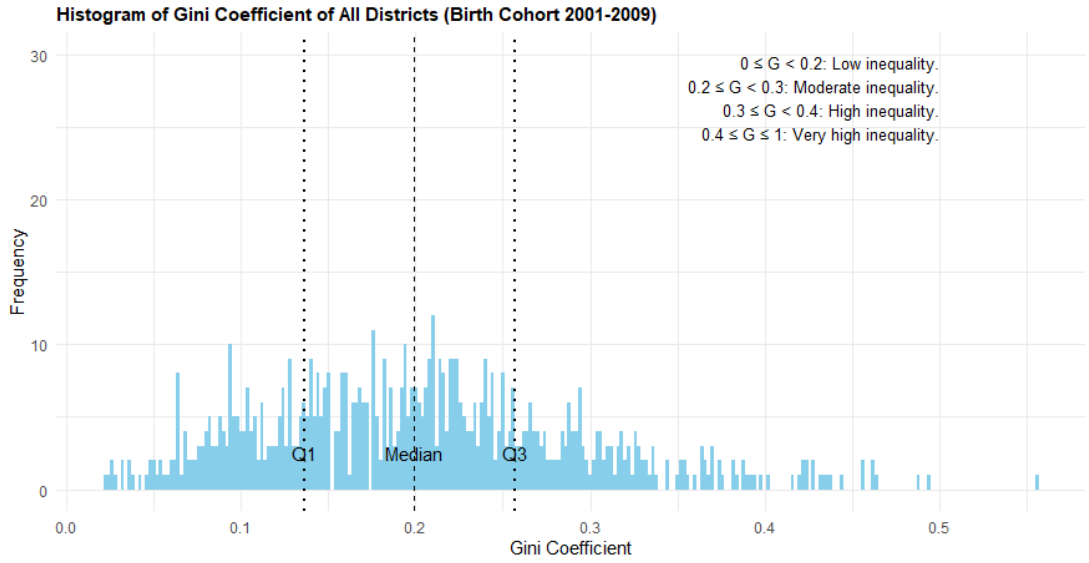


Figure 4



levels of inequality. The QQ-plot of p-values is informative as all 707 p-values (on the y-axis), comparing the Gini coefficients of two birth cohorts across districts, are exceptionally small. In the absence of any pattern of upward mobility, we would expect the p-values to follow a U[0,1] distribution, randomly along the 45-degree line. However, still we observed that even for the birth-cohort 2001-09 the median Gini coefficient was 0.1991259 while the Q3 was 0.2566821 both of which would mean moderate inequality.

Figure 5



This motivates our effort to explore the intersections within this inequality. While there are various methods to approach this, such as Yitzhaki Decomposition (Vakulabharanam 2011), we will adopt a regression model for our analysis.

(4.3) District-Level Regression Results with Bayesian Shrinkage : Birth Cohort 2001-2015

We will reiterate the formula for our within district modelling and state the results, since the results of this set of regressions is the primary findings that we present.

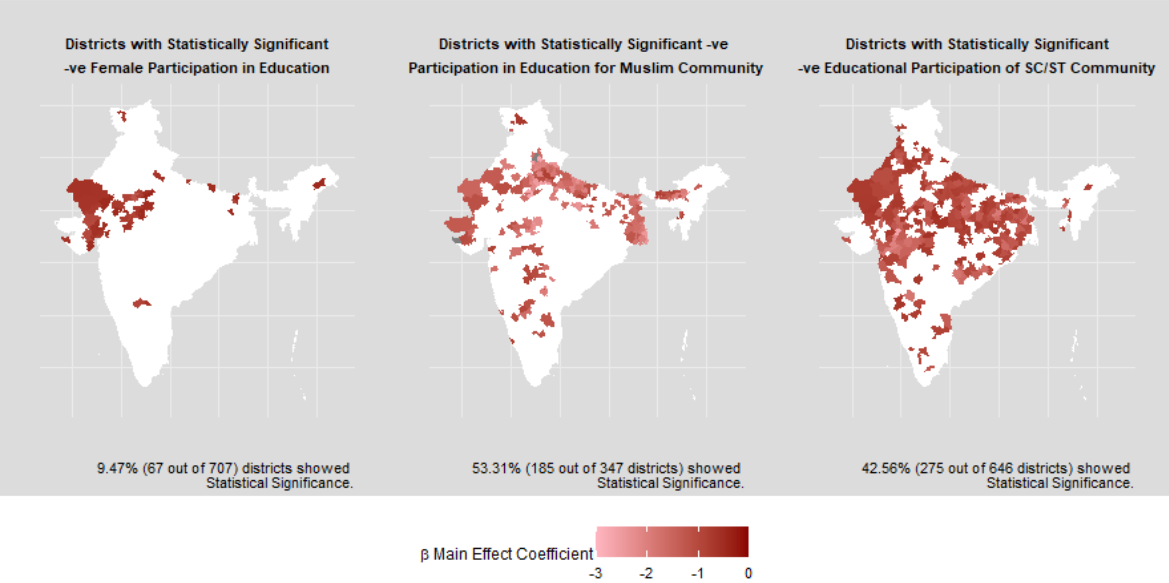
Stage – I

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

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

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

Figure 6



The intercept represents the expected value for forward caste males, serving as the baseline group. We think that mapping the coefficients is the most efficient way to summarize the results. In Figure 6 we display the regression summary results (magnitude and p-values) of the 707 regressions we conducted. In the maps, the main-effect coefficients ($\beta_{female,d}$, $\beta_{SC/ST,d}$ and $\beta_{muslim,d}$) for districts d= 1, 2… 707 are provided for each district where the coefficients were statistically significant. $\beta_{muslim,d}$ indicates how the educational attainment for Muslim males differs from forward caste males. Similarly, $\beta_{SC/ST,d}$ reflects the difference between SC/ST males and forward caste males. $\beta_{female,d}$, captures the difference in the outcome between forward caste females and forward caste males.

The interaction term for Muslim and female measures the additional effect on the outcome for Muslim females beyond the combined individual effects of being Muslim and being female. Likewise, the interaction term for SC/ST and female represents the additional effect for SC/ST females beyond the sum of the effects of being SC/ST and being female. Each term contributes to adjusting the baseline prediction to reflect the outcomes for specific gender and community combinations. However, we did not find any noticeable significance for the interaction terms and that suggested in Muslim and SC/ST households are not any particularly regressive toward women, at least concerning barriers to educational participation in India, based on our 2019-21 data.

The findings clearly demonstrate that, even after accounting for a range of household characteristics, caste and religion play a more significant role than gender in determining educational participation. On average, younger individuals (age 6 to 20) from SC/ST and Muslim communities tend to have 1 to 3 fewer years of education compared to their forward caste counterparts within 42.56 and 53.31 % districts.

While other within-community factors, such as wealth, may contribute, the influence of community remains undeniable even in contemporary India. We emphasize a) consistent with the spatial correlations we observed, it is evident that within certain states, districts may not exhibit a significant community effect. Previous studies, particularly in states like Jammu and Kashmir—where the population is predominantly Muslim—found the community effect to be statistically insignificant. This is unsurprising given the lack of demographic diversity in such states. When studies test community effect for a selected number of states they need at least consider states where there is adequate variance in the data wrt the cofactors being tested b) Moreover, state-level analyses tend to cancel out significant and non-significant results. Certainly, a nationwide single regression fails to capture localized disparities as well . Our study corroborated other studies that conducted granular village level or district level analysis. We stress that the community effect is regionally uneven. c) Furthermore, to ensure accurate regression outcomes, it is crucial to account for multiple records from the same household by incorporating a household-level random effect. When multiple observations (e.g., individuals) come from the same household, their outcomes are likely to be correlated due to shared characteristics (e.g., socioeconomic status, parental education, household resources). If intra-household correlation is ignored, the model treats each individual as an independent data point, effectively inflating the sample size. In reality, the true "effective" sample size is smaller because individuals from the same household do not provide entirely independent information. With inflated sample size, the variance of parameter estimates is underestimated. Since the standard error is derived from this variance, it also becomes underestimated. Underestimated standard errors make test statistics (e.g., t-values) artificially large, leading to p-values that are smaller than they should be. This increases the likelihood of falsely declaring effects statistically significant (Type I error). Confidence intervals become narrower, giving a false impression of precision in the parameter estimates.

(4.4) Quartile Wealth Transition Matrix and Markov Steady-State Probabilities

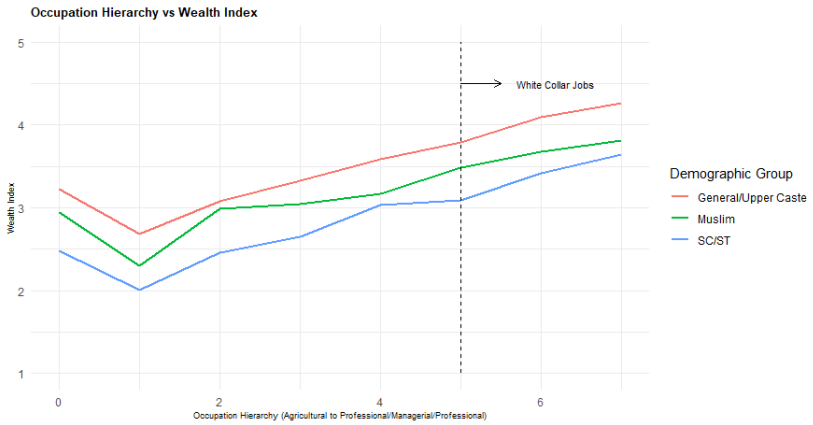
To set the context, we begin with a summary of the national-level occupational hierarchy statistics.

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

Urban Rural	Gender	Education Group	Demographic Group	Not Working	Professional or Technical or Managerial	Clerical	Sales	Services/ Household or Domestic	Agricultural	Skilled/ Unskilled Manual	Other
Urban	Male	Higher	General/Upper Caste	20.5	28.8	5.4	16.1	10.21	3.06	11.36	4.32

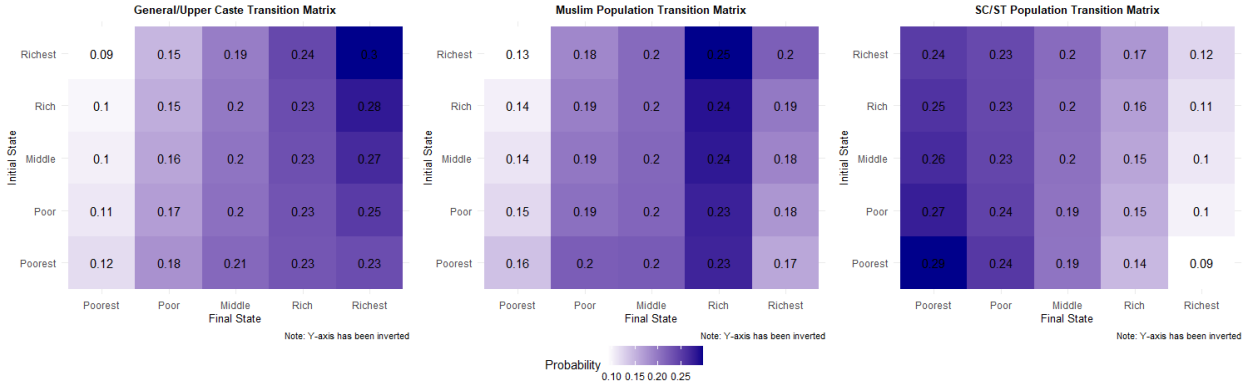
			Muslim	21.9	20.8	4.8	21.2	10.16	3.32	13.86	3.69
			SC/ST	26.7	20.0	6.0	14.5	11.52	3.01	14.12	3.94
		Primary, Secondary	General/Upper Caste	7.8	5.1	2.4	19.5	12.01	9.33	37.78	5.83
			Muslim	7.6	3.2	1.4	19.1	10.30	6.09	43.77	8.33
			SC/ST	9.0	4.2	2.2	11.8	12.98	9.95	43.41	6.30
		No education, Preschool	General/Upper Caste	3.6	1.4	0	14.5	11.67	15.81	47.44	5.30
			Muslim	2.9	0.4	0.9	14.2	8.80	12.18	52.59	7.90
			SC/ST	5.10	1.07	1.6	6.72	12.903	15.86	50.80	5.91
	Female	Higher	General/Upper Caste	95.62	2.54	0.2	0.36	0.52	0.08	0.39	0.17
			Muslim	96.72	1.93	0.1	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	95.9	0.38	0.07	0.66	0.80	0.48	1.43	0.26
			Muslim	97.5	0.12	0.02	0.35	0.46	0.33	0.93	0.16
			SC/ST	95.38	0.29	0.09	0.47	1.20	0.70	1.47	0.38
		No education, Preschool	General/Upper Caste	94.34	0.03	0.04	0.54	1.30	1.35	2.19	0.21
			Muslim	97.03	0.02	0.02	0.28	0.66	0.51	1.24	0.24
			SC/ST	93.30	0.08	0.08	0.41	1.49	1.91	2.30	0.44
Rural	Male	Higher	General/Upper Caste	26	15.43	3.49	8.51	5.21	28.33	9.49	3.55
			Muslim	30	20.76	2.88	11.36	8.64	12.42	10.61	3.33
			SC/ST	29.36	14.21	2.87	5.90	6.10	25.86	12.60	3.07
		Primary, Secondary	General/Upper Caste	7.21	2.05	1.18	7.33	5.47	49.54	23.79	3.43
			Muslim	7.71	2.02	1.22	11.80	6.46	31.47	33.6	5.69
			SC/ST	7.54	1.62	0.93	4.89	5.07	47.46	28.44	4.05
		No education, Preschool	General/Upper Caste	3.64	0.12	0.12	2.71	3.18	65.28	22.48	2.48
			Muslim	3.41	0.69	0	5.59	4.54	45.50	35.72	4.54
			SC/ST	2.85	0.26	0.22	2.30	2.85	63.41	24.63	3.47
	Female	Higher	General/Upper Caste	96.36	1.67	0.12	0.14	0.38	0.74	0.4	0.17
			Muslim	96.89	1.36	0.07	0.07	0.4	0.25	0.83	0.11
			SC/ST	95.62	1.36	0.18	0.25	0.49	1.31	0.56	0.23
		Primary, Secondary	General/Upper Caste	95.35	0.17	0.06	0.25	0.36	2.82	0.80	0.18
			Muslim	97.23	0.13	0.04	0.18	0.44	0.97	0.82	0.19
			SC/ST	94.27	0.23	0.06	0.23	0.50	3.55	0.91	0.26
		No education, Preschool	General/Upper Caste	93.82	0.03	0.02	0.14	0.25	4.83	0.77	0.13
			Muslim	96.51	0.01	0.07	0.18	0.39	1.88	0.76	0.2
			SC/ST	92.73	0.04	0.02	0.16	0.29	5.56	0.95	0.22

Figure 7



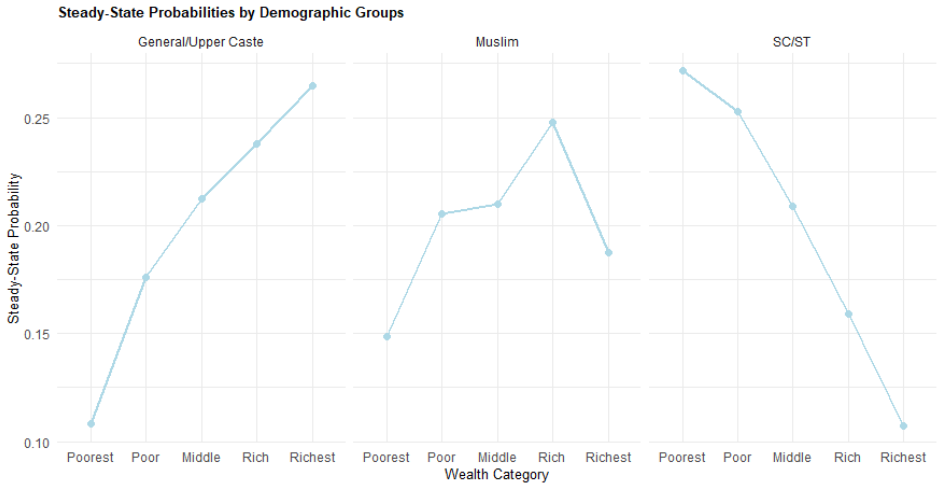
In Figure 7 Wealth Index is Wealth Quartile in the NFHS data. It illustrates the disparities at every level between the general/upper caste, SC/ST, and Muslim communities in terms of their societal standing. In this section, we will explore how seemingly minor differences in occupational proportion (Table 3) can have significant long-term impacts. For SC/ST, Muslim and General/Upper Caste Category the three Transition Matrices we get are :

Figure 8



These are comparable to the quartile education transition matrices presented in previous studies (Maitra and Sharma, 2009; Hnatkovskay et al., 2013). A heatmap makes it much clearer whether communities are catching up or not. The General/Upper Caste group shows significant upward mobility, with a 23% probability of moving from "Middle" to "Rich" and 28% remaining in the "Richest" state.

Figure 9



The Muslim group has lower mobility, with 24% moving from "Poor" to "Middle" and 22% remaining in the "Middle" state. The SC/ST group faces the most challenges, with 28% staying in the "Poorest" state and a 24% probability of falling from "Richest" to "Poorest," highlighting significant barriers to upward mobility.

5. Conclusion

This study identifies critical insights into educational disparities in India through a district-level approach. Spatial clusters of high and low educational attainment, cutting across state boundaries, reveal substantial intra-state heterogeneity. Districts with strong educational outcomes often influence neighboring regions through spillover effects, while isolated districts remain unaffected. These patterns challenge state-level analyses and highlight the need for more spatially informed policy approaches.

Educational inequality has declined significantly across cohorts (2001–2009 vs. 1950–1981), with Gini coefficients reducing across all districts (p-values < 0.001). Despite this, moderate inequality persists, with a median Gini coefficient of 0.20 for the younger cohort, emphasizing the need for targeted interventions.

District-level regressions for 707 districts using mixed-effects models revealed persistent caste and religion effects. SC/ST individuals experience, on average, 1.5 fewer years of education, and Muslim individuals 2.1 fewer years, compared to forward-caste individuals in over 50% of districts. Bayesian posterior adjustments using spatial clusters as priors reduced bias and enhanced statistical power. These regressions also identified substantial variation in community effects across districts, which are obscured in state-level analyses. Interaction terms for Muslim and SC/ST women indicated no significant additional barriers beyond those attributable to community and gender individually.

Wealth mobility analysis reveals caste-based disparities, with SC/ST groups facing the lowest upward mobility and highest persistence in poverty. Transition matrices further illustrate structural barriers to wealth accumulation, supporting the hypothesis that wealth determinants may be endogenous and deeply intertwined with caste effects. These findings underscore the importance of localized analyses and structural reforms to address educational and wealth inequalities in India.