Linguistic Diversity and Development at a Finer Geographic Scale: A Study of Indian Districts

Saarthak Anand[*†]

March 2026

Preliminary draft. Do not cite.

Abstract

This paper examines the relationship between linguistic diversity and economic development using district-level data from India. We construct novel measures of linguistic diversity that incorporate phylogenetic distances between 121 languages recorded in the Indian Census, improving on conventional fractionalisation indices. Using growth in per capita nighttime lights between 2001 and 2011 as a proxy for local economic activity, we document a robust positive and nonlinear relationship between linguistic diversity and economic growth: both the Greenberg and B-indices demonstrate a U-shaped association with the outcome. These results are robust to extensive controls, including geographic characteristics, demographic composition, initial development levels, and state fixed effects, as well as to the exclusion of highly developed districts. To address concerns about unobserved spatial heterogeneity, we additionally employ a border-pair design comparing adjacent districts with large discontinuities in linguistic diversity and find consistent results. Overall, the evidence suggests that the economic benefits of linguistic heterogeneity outweigh potential coordination costs at the local level, with important implications for understanding development in multilingual societies and for policies aimed at preserving linguistic heritage.

Keywords: Linguistic distance, Diversity, Economic development, Greenberg Index, Economic Development

JEL Classification: O15, R11, Z13

[*]Indian Statistical Institute, New Delhi, saarthak21r@isid.ac.in

[†]I am grateful to Dr. Mudit Kapoor for his guidance and supervision. I also thank conference audiences at the Annual Economics Conference at Ahmedabad University, the 2nd Meeting of Young Minds in Frontiers of Economics at IIT Bombay, the 2026 CSW Asia Meeting (Main) at NYU Abu Dhabi, the Research Scholars' Workshop at the University of Calcutta, and Research Scholars' Day at IIT Kanpur for helpful comments and suggestions. All errors are my own.

1 Introduction

The association between cultural diversity and economic development remains one of the most contested questions in development economics. While diversity can generate friction through increased communication costs and social fragmentation, it may also foster innovation and productivity growth by bringing together complementary skills and knowledge systems. This tension is particularly pronounced in linguistically diverse societies, where language differences can simultaneously create barriers to economic exchange and serve as repositories of distinct human capital and cultural practices.

India provides an ideal setting to examine this relationship. With over 700 languages spoken across its territory and 121 languages recognised as mother tongues in the national census, India exhibits extraordinary linguistic diversity both across and within administrative boundaries. This variation, combined with the country's federal structure and diverse economic geography, offers a unique opportunity to identify the causal effects of linguistic diversity on economic outcomes while controlling for other sources of heterogeneity.

The theoretical literature presents competing predictions about diversity's economic effects. Models emphasising communication costs and social cohesion suggest that linguistic fractionalisation impedes economic coordination, increases transaction costs, and undermines the provision of public goods (Easterly and Levine, 1997; Alesina et al., 1999). Conversely, theories highlighting complementarity and knowledge spillovers predict that diversity enhances productivity by facilitating the recombination of different ideas, skills, and cultural practices ((Ottaviano and Peri, 2005).

This paper contributes to this literature by providing the first comprehensive analysis of linguistic diversity and economic growth at the district level in India. I exploit rich data from the 2011 Census of India, which records mother tongue information for 121 languages across all districts and sub-districts, combined with nighttime lights data from the Defense Meteorological Satellite Program as a proxy for economic activity. This approach offers several advantages over existing studies. First, the focus on linguistic rather than ethnic diversity

provides a more precise measure of communication barriers and cultural differences that directly affect economic interactions. Second, the district-level analysis captures economically meaningful geographic units while maintaining sufficient variation for identification. Third, the comprehensive language coverage avoids the selection bias inherent in studies that focus only on major languages or population groups.

A key methodological contribution is the construction of a novel linguistic distance matrix for all 121×121 possible language pairs using phylogenetic data from Ethnologue. This represents a substantial improvement over existing measures that rely on crude categorical distinctions or focus only on major languages. By incorporating the evolutionary relationships between languages, this measure captures the degree to which linguistic differences translate into communication barriers and cultural distance.

The empirical analysis indicates a robust positive relationship between linguistic diversity and economic development across Indian districts. Using per capita nighttime lights growth between 2001 and 2011 as the outcome, the preferred specification with state fixed effects (Model 4) shows that both the Greenberg and B-indices of linguistic diversity have a positive and statistically significant association with growth, with their squared terms also significant, suggesting a nonlinear, convex relationship. The magnitude implies that districts with higher initial linguistic diversity experienced faster growth in nighttime lights, even after controlling for demographic, geographic, and initial development characteristics. These results are robust when using an alternative measure of diversity, as well as when excluding the top 5% or 10% of districts by 2011 nighttime lights intensity, indicating that the findings are not driven by outliers or highly urbanised districts. Including the logarithm of distance to the state capital in the B-index specification slightly reduces the coefficient but does not alter the significance, further supporting the robustness of the results. Overall, the magnitudes of the diversity coefficients are comparable across indices, highlighting that linguistic diversity consistently exhibits a positive effect on district-level economic outcomes, aligning with recent studies that find positive associations between cultural or linguistic diversity

and development at more granular geographic levels (Montalvo and Reynal-Querol, 2021; Baskaran et al., 2023).

These findings have important implications for understanding the microfoundations of diversity's economic effects and for policy debates about language preservation and cultural assimilation in developing countries. By demonstrating positive effects of linguistic diversity at the district level, this study contributes to a growing body of evidence that challenges pessimistic views about diversity's economic consequences and highlights the potential value of cultural preservation in promoting inclusive growth.

2 Related Literature

The association between diversity and economic development has generated extensive empirical investigation, with findings that vary significantly across geographic scales and institutional contexts. Early cross-country studies predominantly documented negative relationships between ethnic diversity and economic growth, establishing diversity as an important determinant of development outcomes.

2.1 Cross-Country Evidence on Diversity and Growth

Easterly and Levine (1997) examine the effect of ethnic diversity on economic growth both directly and indirectly through its impact on public policies. They argue that the literature inadequately addresses why some nations opt for growth-enhancing policies while others choose growth-retarding ones. Their findings suggest that a shift from complete homogeneity to complete heterogeneity is associated with a fall of two percentage points in economic growth. This seminal work established diversity as a key determinant of economic outcomes and influenced subsequent research in the field.

A crucial mechanism linking diversity to economic development operates through public goods provision. Alesina et al. (1999) demonstrate that city-provided public goods such as

roads and education are inversely related to ethnic fragmentation. In ethnically diverse cities, politicians often serve constituencies based on ethnicity and are likely to value the benefits of public goods accruing to their own ethnic voter base while discounting those accruing to other groups. This political economy channel provides a concrete explanation for diversity's negative effects on development outcomes.

2.2 The Geographic Scale of Diversity Effects

Recent empirical evidence has begun to reconcile competing theoretical perspectives by demonstrating that diversity's net effects depend critically on the level of analysis and institutional context. While cross-country studies often find negative associations between ethnic fractionalisation and growth, micro-level analyses increasingly document positive effects of diversity on innovation, productivity, and economic performance in cities and regions with strong institutions. In the Indian context, Jain (2017) provides causal evidence that linguistic mismatch between the official language of instruction and the mother tongue significantly reduced literacy and college attainment in colonial districts, with effects persisting over time. Exploiting the exogenous formation of colonial provinces and the 1956 linguistic reorganisation of states, the study highlights how language barriers in education shape long-run human capital accumulation. This evidence underscores the importance of communication costs at the local level and motivates a broader examination of how linguistic structure within districts relates to economic performance.

Montalvo and Reynal-Querol (2021) examine the relationship between ethnic heterogeneity and economic growth at the city level, finding a positive association between diversity and wages and productivity. In the African context, this result is driven by increased trade at the boundaries between various ethnic groups due to specialisation at the ethnic group level. This suggests that the benefits of diversity may be more readily realised at smaller geographic scales where coordination costs are lower and agglomeration economies can amplify positive spillovers.

2.3 Linguistic Distance and Cultural Proximity

This paper contributes to the growing literature on linguistic distance as a measure of cultural proximity. Fearon (2003) constructs a measure of ethnic fractionalisation based on 822 ethnic groups in 160 countries, with an additional index incorporating linguistic distance based on common tree branches between languages according to *Ethnologue* as a proxy for cultural distance between groups.

Desmet et al. (2005) examine the effect of ethnolinguistic conflict on redistribution using an index of linguistic conflict that incorporates linguistic distance based on language family information from *Ethnologue*. However, their approach differs by not utilising language trees. They find that incorporating linguistic distances makes both linguistic diversity and polarisation better predictors of redistribution. Desmet et al. (2009) employs tree diagrams and demonstrates that while commonly used fractionalisation measures yield insignificant results in cross-country analysis, accounting for linguistic distances makes linguistic diversity have both statistically and economically significant impacts on redistribution.

The basic fractionalisation index used in this literature was developed independently across multiple disciplines. In linguistics, it was introduced by Greenberg (1956) as the *monolingual non-weighted index* or A-index, providing the theoretical foundation for contemporary diversity measures.

2.4 Nighttime Lights as Economic Activity Proxy

Recent economic research has utilised nighttime lights data for measuring economic activity, particularly in contexts where traditional economic statistics are unavailable or unreliable. Elvidge et al. (2007) demonstrate that nighttime light intensity is highly correlated with gross domestic product and electric power consumption. Henderson et al. (2012) show that satellite nighttime lights data serve as useful proxies for economic activity when traditional data are unavailable or of poor quality, particularly in developing countries at subnational levels such as districts and sub-districts.

Asher et al. (2021a) confirm that nighttime lights are highly significant proxies for population, employment, per capita consumption, and electrification at very local levels. Baskaran et al. (2023) exemplify this approach by using nighttime lights to measure economic performance at the level of electoral constituencies in Indian states to assess the impact of female legislators on economic outcomes. This methodological foundation supports the use of nighttime lights as our primary measure of district-level economic activity.

3 Data

The language data of the lists the number of individuals in every district and subdistrict who identify a particular language as their mother tongue. For a language to be considered a mother tongue in the Census, it must have atleast 10,000 speakers. The 2001 Census identifies 22 Scheduled and 100 Non-Scheduled languages, belonging to 5 different families. Table 1 shows the number and percentage of the total population identifying languages from each of the six languages families as mother tongue. The language family having the greatest number of speakers in India is the Indo-Aryan family.

I use per capita nighttime lights intensity as a proxy for economic activity. Intensity of nighttime lights indicates indoor and outdoor activity. Consumption of nearly all goods in the evening requires lights. As income rises, so does lights usage per person, in both consumption activities and many investment activities. Satellites from the United States Air Force Defense Meteorological Satellite Program (DMSP) have been circling the earth 14 times per day recording the intensity of Earth-based lights with their Operational Linescan System (OLS) sensors since the 1970s, with a digital archive beginning in 1992. Each satellite-year dataset is a grid reporting the intensity of lights as a six-bit digital number, for every 30 arc-second output pixel (approximately 0.86 square kilometers at the equator) between 65 degrees south and 75 degrees north latitude. Value of nighttime light intensity is recorded as a digital number (DN) between 0 and 63. This data is sourced from SHRUG (Asher et al.

(2021a)).

Some of the covariates in our analysis include the literacy rate of the district in the 2011 Census, the sex ratio, the percentage of population living in urban areas, the percentage of population that is Scheduled Caste. Another covariate is the Terrain Ruggedness Index (TRI), constructed from the elevation data The TRI expresses the elevation difference between adjacent pixels. It is the square root of the mean squared difference in elevation values between a pixel and the eight cells immediately surrounding it. Greater ruggedness is associated with a higher cost to development.

3.1 Diversity indices

The Greenberg index is our first measure of linguistic diversity. Introduced to linguistics first by Greenberg (1956), it is the probability that two individuals randomly selected from the population have different mother tongues.

$$A = 1 - \Sigma_i (p_i/P)^2 \tag{1}$$

where p_i is the number of people in the population speaking language i, and P is the total population size of the region. The index ranges from 0 to 1, and more linguistically diverse regions have a relatively higher value. Regions where a single language is the mother tongue have an LDI value of 0, and those where every individual speaks a different mother tongue have an LDI value of 1.

An alternate diversity index is the Shannon index (H) (Shannon (1948)). In a less diverse population, the uncertainty of prediction is low. Events with a certain outcome have no uncertainty, and thus a perfectly homogeneous society has an H value of 0. Unlike the LDI, however, the Shannon index does not have an upper bound of 1.

3.2 Linguistic Distances

The Greenberg and Shannon indices do not take into consideration how different any two languages are from each other. Hindi and Punjabi, for instance, are less different from each other than Hindi and Malayalam. It would thus be relatively easier for a person with Hindi as mother tongue to learn Punjabi than Malayalam, and the expected transaction cost would be lower. For this purpose, I use a concept called linguistic distance which is logically orthogonal to economic outcomes. Linguistic distance is a measure of relative degree of similarity between a pair of languages.

I use the linguistic trees from Ethnologue to calculate linguistic distance. The principle is that of genealogical trees, which go from a shared common ancestor to successive generations. Languages which are similar to each other are likely to have been separated at a later stage, and thus share more number nodes. These languages have a more recent common ancestor. Languages which belong to different families altogether are different from each other, and don't share a common ancestor. The number of common nodes between a pair of languages is thus a measure of linguistic distance between those languages.

Another way of calculating linguistic distance involves lexicostatics, which is a method of comparative linguistics that involves comparing the percentage of lexical cognates between languages to determine their relationship. Dyen et al. (1992) assembled data on 200 common "meanings" from all Indo-European languages. For each language, they compiled lists of words expressing these meanings. When words from two languages originated from a common source, they're classified as "cognate". Aggregating over the 200 meanings, the measure of linguistic distance is the percentage of noncognate words. The linguistic distance between two languages is lower, and the similarity greater, more the number of number of cognate words between them.

Ethnologue contains detailed genealogical information on living languages of the world, including the ones contained in the Census of India. Linguistic trees include all the nodes which lead to a particular language. I use these linguistic trees to calculate linguistic distance.

Linguistic distance between a pair of languages is calculated as the ratio of the number of common nodes between the languages and the maximum number of common nodes between any two languages in our analysis. Formally,

$$d_{ij} = \frac{MaxCN - CN_{ij}}{CN_{ij}} \tag{2}$$

where d_{ij} is the linguistic distance between languages i and j, $MaxCN$ is the maximum number of common between any two languages, CN_{ij} is the number of common nodes between languages i and j. d_{ij} lies between 0 and 1. It decreases with an increase in the number of common nodes between the languages. In our dataset which includes the 121 languages in the 2011 Census of India, the $MaxCN$ is 6. The formula for linguistic distance becomes.

$$d_{ij} = \frac{6 - CN_{ij}}{CN_{ij}} \tag{3}$$

Once I incorporate linguistic distance in our diversity index, the resulting index, which, following Greenberg (1956), I refer to as the B-Index, can be written as,

$$B = \sum_i \sum_j (s_i \times s_j \times d_{ij}) \tag{4}$$

where B is the B-Index. It can be interpreted as the expected linguistic distance between two randomly selected individuals from the population. Similar to the Greenberg Index and the Shannon Index, a higher value indicates greater linguistic diversity or fractionalisation in the region. A region with only one language have a single linguistic distance value of 0, and, consequently, the B-Index value would thus be 0 as well, indicating perfect linguistic homogeneity in the region.

This analysis required manually entering information from Ethnologue on all the nodes that lead up to the 121 languages listed as mother tongues in the 2011 Census. Using this dataset, CN_{ij}, the number of common nodes between all possible language pairs was

calculated. Languages belonging to different families have a CN_{ij} value of 0, since they do not share any nodes.

Figures 1 and 2 present the spatial distribution of linguistic diversity across Indian districts in 2011, measured by the Greenberg Index and B-Index respectively. Both measures reveal substantial geographic heterogeneity in linguistic composition, with the highest diversity concentrated in the northeastern states, parts of central India (particularly tribal areas), and select districts in the Western Ghats and southern peninsular regions. The Greenberg Index exhibits a broader range of valuescompared to the B-Index , reflecting differences in how these measures weight linguistic distances versus simple fractionalisation. Notably, both indices identify similar geographic clusters of high diversity, particularly in states like Assam, Meghalaya, and parts of Odisha and Chhattisgarh, while the Indo-Gangetic plain displays relatively lower linguistic diversity across both measures.

3.3 Summary Statistics

Table 2 presents summary statistics for the key variables employed in our analysis across Indian districts for 2001 and 2011. The data reveal several notable patterns that provide important context for our empirical investigation.

Our outcome, log per capita nighttime lights, shows an improvement in economic activity over the decade, with mean values increasing from -13.06 in 2001 to -12.39 in 2011, accompanied by reduced dispersion (standard deviation declining from 1.19 to 1.04). This suggests both economic growth and some convergence in development levels across districts during this period.

Regarding linguistic diversity, the Greenberg Index exhibits a slight decline in mean diversity from 0.29 to 0.26 between 2001 and 2011, with the median falling more substantially from 0.24 to 0.19. This pattern may reflect ongoing linguistic consolidation processes or changes in population composition due to internal migration. The B-Index, which incorporates linguistic distances between languages, exhibits a similar trajectory.

The covariates demonstrate expected developmental trends. Literacy rates improved substantially over the decade, rising from a mean of 0.54 to 0.63, with reduced dispersion indicating convergence in educational outcomes. The sex ratio remained relatively stable, with a slight improvement in gender balance from 0.48 to 0.49. The proportion of Scheduled Caste population shows minimal variation both across districts and over time, with means of 0.14 and 0.15 respectively.

Geographic controls are naturally constant over the two periods. The Terrain Ruggedness Index and Distance have means of 10.05 and 496.83 kilometers respectively. The high standard deviation for terrain ruggedness (10.90) and distance to coast (385.81) indicates substantial geographic heterogeneity across Indian districts, underscoring the importance of controlling for these spatial factors in our analysis.

The summary statistics reveal considerable variation in both our key variables of interest and control variables, providing sufficient identifying variation for our empirical strategy. The temporal changes observed, particularly in nighttime lights and literacy, align with India's broader developmental trajectory during this period, while the contrasting patterns in the two diversity measures highlight the importance of examining multiple conceptualisations of linguistic heterogeneity.

4 Methodology

4.1 Empirical Specification

To examine the relationship between ethnic diversity and economic growth at the district level, we employ a cross-sectional OLS regression framework using the change in log per capita nighttime lights intensity between 2001 and 2011 as our dependent variable. Nighttime lights data, sourced from the Defense Meteorological Satellite Program (DMSP), serve as a proxy for economic activity and development, following the approach established in the literature (Henderson et al., 2012).

Our baseline empirical specification is given by:

$$\log(\eta_{2011,i}) - \log(\eta_{2001,i}) = \beta_0 + \beta_1 \cdot \text{Diversity}_{2001,i} + \beta_2 \cdot \text{Diversity}^2_{2001,i} + \beta_3 \cdot \log(\eta_{2001,i}) + \beta_4 \cdot x_i + \varepsilon_i \quad (5)$$

where $\eta_{t,i}$ represents per capita nighttime lights intensity in district i at time t. The dependent variable thus captures the growth rate of economic activity over the decade from 2001 to 2011. Our main explanatory variable, $\text{Diversity}_{2001,i}$, is measured using the Greenberg diversity index and the B-Index in the two specifications. They are calculated from district-level demographic data from the 2001 Census.

The quadratic specification allows for non-linear relationships between diversity and growth, enabling us to test whether the diversity-growth relationship exhibits diminishing returns or threshold effects. The inclusion of the initial level of nighttime lights intensity, $\log(\eta_{2001,i})$, controls for convergence effects and ensures our estimates capture growth dynamics rather than level differences.

The vector x_i includes a comprehensive set of district-level control variables measured in 2001 to account for other factors that may influence economic growth. These controls include demographic characteristics (population density, urbanisation rate), educational attainment (literacy rates), infrastructure measures, and geographic features that may affect development outcomes.

4.2 Addressing Spatial Heterogeneity

To account for unobserved state-level heterogeneity that may confound the relationship between diversity and growth, we extend our baseline specification to include state fixed effects:

$$\log(\eta_{2011,i}) - \log(\eta_{2001,i}) = \beta_0 + \beta_1 \cdot \text{Diversity}_{2001,i} + \beta_2 \cdot \text{Diversity}^2_{2001,i} + \beta_3 \cdot \log(\eta_{2001,i}) + \beta_4 \cdot x_i + \lambda_s + \varepsilon_i \quad (6)$$

where λ_s represents state fixed effects. This specification controls for state-level policies, institutional differences, natural resource endowments, and other unobserved characteristics that may simultaneously affect both diversity patterns and economic growth within states. By including these fixed effects, we identify the diversity-growth relationship from within-state variation across districts, providing a more rigorous test of our hypothesis.

4.3 Identification Strategy

Our identification strategy relies on the assumption that, conditional on initial economic conditions and observable district characteristics, the level of ethnic diversity in 2001 is exogenous to subsequent growth outcomes. The inclusion of initial nighttime lights intensity and comprehensive controls helps address concerns about omitted variable bias, while state fixed effects further mitigate the influence of unobserved spatial heterogeneity.

Standard errors are clustered at the state level to account for potential spatial correlation in the error terms and to ensure valid statistical inference in the presence of state fixed effects.

5 Results

Table 3 presents results from a series of OLS regressions examining the association between linguistic diversity and economic growth at the district level. The outcome variable is the change in the log of per capita nighttime lights between 2001 and 2011, a proxy for local economic growth.

Column (1) includes only the Greenberg Index (2001) and its square. *Column (2)* adds

baseline demographic covariates: sex ratio, SC share, literacy rate, and initial log of per capita nighttime lights (2001). *Column (3)* adds geographic controls: terrain ruggedness and the log of distance to the coast. *Column (4)* introduces state fixed effects, accounting for time-invariant state-level characteristics.

In all specifications, the coefficient on the Greenberg Index is negative, while the coefficient on its square is positive and statistically significant. This implies a U-shaped relationship between initial linguistic diversity and subsequent growth: districts with moderate levels of diversity experience slower growth, whereas those with either low or high diversity tend to grow faster. This nonlinear pattern is robust to the inclusion of controls and state fixed effects.

In the most comprehensive specification (Model 4), the Greenberg Index has a coefficient of -0.89 and its square has a coefficient of 1.35, both statistically significant at the 1% level. This provides evidence of a non-monotonic relationship between ethnolinguistic diversity and local development outcomes.

Table 4 includes the B-index, which incorporates linguistic distances, as the predictor. In the preferred specification (Model 4), which includes full demographic and geographic controls along with state fixed effects, we find a statistically significant U-shaped relationship between the B-Index and district-level economic growth. The coefficient on the B-Index is -0.79 and that on its square is 1.53, both significant at the 5% and 1% levels respectively. This implies that the marginal effect of diversity on growth is initially negative but turns positive beyond a certain threshold.

The turning point of this relationship occurs at a B-Index value of approximately 0.26, calculated by taking the ratio $\frac{0.79}{2\times 1.53}$. Below this threshold, increasing diversity is associated with lower growth; above it, greater diversity is positively associated with growth. This non-monotonic relationship echoes the pattern observed with the Greenberg Index, providing consistent evidence that the effect of ethnolinguistic diversity on development is highly nonlinear.

Comparing the results across the two diversity measures, we find that the magnitudes of the estimated effects are broadly similar in the preferred specifications. In Model 4 of Table 3, the coefficient on the Greenberg Index is -0.89, with the square term at 1.35. In Model 4 of Table 4, the corresponding coefficients for the B-Index are -0.79 and 1.53, respectively.

While the negative linear terms are of comparable size, the curvature implied by the square term is slightly steeper in the B-Index case, suggesting a somewhat stronger U-shaped relationship. However, both sets of estimates indicate a robust and consistent pattern: linguistic diversity tends to have a nonlinear relationship with growth, with low or high levels of diversity being more conducive to growth than intermediate levels. This similarity in both sign and magnitude across indices lends confidence to the substantive interpretation of the results.

6 Robustness Checks

To ensure that our main findings are not driven by specific modeling choices or influential observations, we conduct a series of robustness checks. These exercises examine the sensitivity of the estimated relationship between linguistic diversity and economic development to alternative specifications, sample restrictions, and additional control variables. The results from these checks provide further confidence in the validity and reliability of our primary analysis.

6.1 Distance to administrative centres

As a robustness check, we augment our baseline specification with the logarithm of distance to the state capital as an additional control variable. This geographic control addresses potential confounding arising from systematic spatial variation in both economic development and linguistic composition across districts within states. Distance to state capitals proxies for differential access to administrative services, concentrated government expenditures,

and agglomeration economies that typically diminish with distance from major urban centers (Henderson, 2003; Krugman, 1991). Importantly, linguistic diversity may exhibit non-random spatial patterns due to historical migration flows, colonial administrative boundaries, or the geographic distribution of indigenous populations, potentially correlating with proximity to political centers (**?**). The logarithmic transformation captures the well-documented non-linear relationship between distance and economic activity, wherein the marginal effect of distance decays at greater distances from urban centers (Redding and Rossi-Hansberg, 2017). By controlling for this spatial dimension, we ensure that our estimated association between linguistic diversity and economic development is not confounded by location-specific advantages or disadvantages relative to state-level administrative and economic hubs.

In *column (5)* of Table‘4, we introduce the logarithm of the distance to the state capital as a robustness check, to account for potential spatial variation in access to administrative or economic hubs. The core relationship between the B-Index and district-level growth remains largely intact: the linear term remains negative (-0.84), and the squared term remains positive and statistically significant (1.72). The implied turning point is at a B-Index value of approximately 0.24, very close to that in the baseline specification, suggesting that the observed U-shaped relationship is not sensitive to the inclusion of this geographic control.

Interestingly, the coefficient on log distance to the capital itself is small and statistically insignificant, indicating that once demographic and geographic controls are included, remoteness from the capital city does not exert an independent influence on nighttime lights growth. This enhances confidence that the observed diversity-growth relationship is not merely reflecting urban-proximity effects.

6.2 Alternate measure of fractionalisation

To ensure our results are not sensitive to the specific measure of linguistic diversity employed, I re-estimate the baseline specification using the Shannon index in place of the Greenberg index. While both indices capture linguistic fractionalisation, they differ in their mathemat-

ical properties and sensitivity to the distribution of linguistic groups (Shannon, 1948). The Shannon index, derived from information theory, places greater weight on the evenness of the linguistic distribution and is more sensitive to the presence of numerous small linguistic groups, whereas the Greenberg index emphasises the probability that two randomly selected individuals speak different languages. Developed from information theory, it is a measure of entropy in a population.

$$H = -\Sigma_i(p_i * ln(p_i)) \tag{7}$$

This alternative specification serves as a crucial robustness check, as different diversity measures may capture distinct aspects of linguistic heterogeneity that could have varying economic implications (**?**Fearon and Laitin, 2003). Consistency of results across these alternative measures would strengthen confidence in our findings and suggest that the observed associations are not artifacts of our particular choice of diversity metric.

Table 5 reports the regression results using the Shannon Index as an alternative measure of linguistic diversity. Across specifications, we observe a consistent and statistically significant non-linear (U-shaped) relationship between the Shannon Index in 2001 and subsequent district-level growth in per capita nighttime lights. Model 4, which includes demographic and geographic controls along with state fixed effects, serves as our preferred specification in this case too.

In Model 4, the coefficient on the linear term of the Shannon Index is negative (-0.79), while the squared term is positive and highly significant (1.53). This mirrors the pattern observed with the Greenberg Index and the B-Index, suggesting that both measures of linguistic diversity capture a similar underlying nonlinear relationship with economic growth. The negative coefficient on the linear term indicates that, at low levels of diversity, increases in diversity are associated with slower growth. However, the positive and convex nature of the squared term implies that beyond a certain threshold, additional diversity is associated with higher growth.

To interpret the magnitude of the effect, the implied turning point of the U-shaped curve is approximately 0.26 on the Shannon Index scale. This means that districts with a Shannon Index below this threshold experience a negative relationship between diversity and growth, while those above the threshold benefit from increased diversity. The overall shape and significance of the relationship are comparable in size and direction to the Greenberg-based results, reinforcing the robustness of the findings to alternative measures of linguistic diversity.

6.3 Agglomeration Effects

To ensure our results are not driven by urban agglomeration economies or outlier observations, we conduct a robustness check that excludes the top 5% and 10% of districts ranked by per capita nighttime lights intensity in 2011. This approach addresses concerns that our findings may be primarily driven by highly urbanised districts where agglomeration effects, rather than diversity per se, generate economic growth (Henderson, 2003). Urban centers typically exhibit both higher diversity and faster growth due to economies of scale, knowledge spillovers, and market access advantages that are distinct from diversity-driven mechanisms. By excluding these top-performing districts, we can assess whether the diversity-growth relationship holds in the broader sample of less urbanised areas where agglomeration effects are likely to be less pronounced.

To account for the possibility that the main results are driven by a small number of highly urbanised or economically advanced districts, we conduct robustness checks by excluding the top 5% and 10% of districts based on per capita nighttime lights intensity in 2011. Table 6 presents the results using Model 4, our preferred specification, with the B-Index as the measure of linguistic diversity.

The core findings remain robust to these exclusions. In both versions of the trimmed sample, the coefficient on the B-Index remains negative, and its square remains positive and statistically significant. The magnitudes of the coefficients are only modestly attenuated, with the

linear term ranging between -0.76 and -0.81, and the squared term between 1.49 and 1.62. These results reaffirm the non-linear, U-shaped association between linguistic diversity and district-level growth.

The persistence of this pattern suggests that the observed relationship is not being driven by agglomeration effects specific to the most developed districts. Instead, the U-shaped link between diversity and growth holds more generally across the urban-rural spectrum.

7 Border-Pair Research Design

Estimating the relationship between linguistic diversity and economic growth is complicated by the fact that linguistic composition is correlated with geography, historical settlement patterns, and institutions. To mitigate these concerns, we employ a border-pair research design that exploits sharp discontinuities in linguistic diversity across administrative district boundaries.

The unit of analysis is a district-level grid cell. We identify pairs of geographically adjacent administrative districts that share a common boundary and construct border-pairs from these neighboring districts. Importantly, administrative borders are not defined on the basis of language. Instead, linguistic diversity varies within and across districts.

To sharpen identification, we restrict the sample to border-pairs in the top 10 percent of the distribution of the absolute difference in the linguistic B-Index across all district border-pairs. This restriction ensures that the analysis focuses on neighboring districts that exhibit large discontinuities in linguistic composition while remaining geographically proximate.

Figure 3 illustrates the construction of the border-pairs. For each administrative boundary, we retain districts within a fixed bandwidth and compare outcomes only within matched pairs that display large differences in linguistic diversity.

We estimate regressions of the form:

$$\Delta \log y_{ip} = \beta \, \text{Linguistic Diversity}_{ip} + X'_{ip}\gamma + \mu_p + \varepsilon_{ip}, \tag{8}$$

where i indexes districts and p indexes border-pairs. The inclusion of border-pair fixed effects μ_p ensures that identification comes exclusively from comparisons between adjacent districts on opposite sides of the same administrative boundary.

Table 7 presents the baseline results. Across specifications, districts with higher linguistic diversity experience faster subsequent growth. The estimates are stable to the inclusion of geographic controls and alternative bandwidths.

The identifying assumption underlying this design is that unobserved determinants of economic growth vary smoothly across administrative borders. Under this assumption, differences in growth across paired districts can be attributed to differences in linguistic diversity rather than to geography or institutions. The estimates should be interpreted as local effects identified at administrative boundaries with large discontinuities in linguistic composition.

8 Mechanism: Linguistic Diversity and Market Thickness

This section examines a potential mechanism linking linguistic diversity to economic development: market thickness. A large literature in urban and spatial economics emphasises that local economic performance depends critically on the depth of markets, as thicker markets improve matching between workers and firms, enable greater specialisation, and support production at scale (Krugman, 1991; Henderson, 2003; Redding and Rossi-Hansberg, 2017). These agglomeration forces are typically strongest at fine geographic scales, where repeated interaction and local institutions can mitigate coordination frictions.

Linguistic diversity may contribute to market thickness by expanding the effective pool

of skills, occupations, and consumer demand within a district. When individuals from different linguistic backgrounds interact in local labour and product markets, the scope for complementary matching increases. This mechanism is closely related to work highlighting the productive benefits of cultural and linguistic diversity in cities and regions (Ottaviano and Peri, 2005; Montalvo and Reynal-Querol, 2021), and to theories of agglomeration that emphasise diversity-driven complementarities (Duranton and Puga, 2004).

8.1 Conceptual Framework

Linguistic heterogeneity can affect local economic outcomes through two opposing channels. On the one hand, language differences may raise communication and coordination costs, potentially fragmenting markets and reducing scale. On the other hand, linguistic diversity may enhance complementarities by bringing together heterogeneous skills, knowledge, and preferences, thereby deepening markets and improving matching efficiency.

At the district level, where economic interaction is spatially concentrated and repeated, the second channel may dominate. In such settings, linguistic diversity can support a wider range of services and occupations, increase the scale of local demand, and reduce effective fixed costs faced by firms. These forces should manifest empirically as denser labour markets, higher establishment growth, and a shift in firm structure towards larger establishments.

8.2 Data and Empirical Approach

To investigate this mechanism, I use district-level data from the Economic Census of India, which provides comprehensive coverage of non-farm establishments across manufacturing and services. I draw on two waves of the census, 2005 and 2013, covering both formal and informal firms, including own-account enterprises and larger establishments.

I operationalise market thickness using three sets of outcomes. First, I examine non-farm employment density, measured as log non-farm employment relative to population, which captures the concentration of workers and jobs in the local economy. Second, I analyse growth

in the number of establishments between 2005 and 2013 as a measure of entry and expansion. Third, I study firm size composition, focusing on the share of very small establishments (fewer than 20 workers) and the share of large establishments (100 or more workers), which proxy for the viability of scale and specialisation (Ghani et al., 2014).

The empirical specifications relate these outcomes to baseline linguistic diversity, controlling for initial development, demographic characteristics, and state fixed effects. As with much of the agglomeration literature, these estimates are interpreted as suggestive evidence on mechanisms rather than as causal effects.

8.3 Results

The results, given in Table 8, are consistent with a market-thickness channel. Districts with higher linguistic diversity exhibit significantly higher non-farm employment density, indicating deeper local labour markets. Linguistic diversity is also associated with faster growth in the number of establishments between 2005 and 2013, suggesting greater entry or expansion of economic activity.

In terms of firm structure, linguistically diverse districts display a systematic shift away from very small establishments towards larger firms. The share of establishments with fewer than 20 workers is lower, while the share of establishments with 100 or more workers is higher, in districts with greater linguistic diversity. These patterns indicate that more linguistically heterogeneous districts are better able to sustain scale and specialisation.

8.4 Interpretation

These findings suggest that linguistic diversity may facilitate local economic development by deepening markets rather than fragmenting them. While linguistic heterogeneity may raise coordination costs in some settings, at the district level these costs appear to be outweighed by gains from matching, specialisation, and scale emphasised in the agglomeration literature.

Importantly, this evidence does not identify a single causal pathway. Instead, it provides

support for a plausible mechanism through which linguistic diversity can be positively associated with growth at fine geographic scales, complementing the baseline and border-pair results.

9 Conclusion and Discussion

This paper provides a comprehensive analysis of the relationship between linguistic diversity and economic development at the district level in India. By combining detailed mother tongue data from the Census of India with satellite-based nighttime lights as a proxy for economic activity, we exploit variation across 619 districts to identify the effects of linguistic diversity on local growth outcomes.

The results consistently show a positive association between linguistic diversity and economic development. In the preferred specifications with state fixed effects, higher initial diversity—as measured by the various indices—is associated with faster growth in per capita nighttime lights between 2001 and 2011. The significance of both the linear and squared terms suggests a nonlinear, convex relationship, implying that the benefits of diversity may increase up to a point but taper off in extremely heterogeneous districts. These findings are robust to alternative measures of diversity, the exclusion of highly developed districts, and the inclusion of additional geographic controls such as distance to the state capital.

To further strengthen the empirical interpretation, the paper employs a border-pair research design that compares adjacent districts on opposite sides of administrative boundaries with large discontinuities in linguistic diversity. Focusing on district pairs in the top decile of differences in the linguistic B-Index, this design mitigates concerns related to unobserved geographic or institutional heterogeneity. The persistence of positive effects in this restricted sample reinforces the view that linguistic diversity is not merely correlated with development, but operates through economically meaningful local variation.

The empirical evidence contributes to the broader literature on diversity and develop-

ment by demonstrating that the positive effects of linguistic heterogeneity persist even at fine geographic levels. Unlike studies that focus on national- or state-level aggregates, this district-level analysis captures economically meaningful spatial variation, highlighting the potential of diversity to foster knowledge spillovers, innovation, and complementary skill accumulation at the local scale.

While the analysis is not designed to pin down a single mechanism, the findings are consistent with channels operating through local interaction across language groups, including labor market complementarities, migration and sorting, and improved market integration in linguistically diverse areas. These mechanisms are likely to be especially relevant in contexts such as India, where multilingualism is widespread and economic exchange frequently spans linguistic boundaries.

Policy implications from these results are twofold. First, they suggest that cultural and linguistic diversity should not be viewed solely as a source of coordination challenges; rather, diversity can serve as an asset that enhances productivity when combined with effective governance and inclusive institutions. Second, policies that promote the preservation of linguistic heritage and encourage intercultural communication could further strengthen the economic benefits of diversity, particularly in regions with high linguistic heterogeneity.

Future research could extend this analysis by exploring the mechanisms through which linguistic diversity translates into local economic growth, such as the role of migration, trade, or knowledge networks. Moreover, integrating household- or firm-level data could shed light on micro-level channels, while longitudinal analyses spanning additional decades could examine the persistence of these effects over time. Overall, this study underscores the nuanced role of linguistic diversity in shaping economic development and provides evidence supporting the potential benefits of embracing cultural heterogeneity in policy and planning.

References

Alesina, A., R. Baqir, and W. Easterly (1999): "Public goods and ethnic divisions," *The Quarterly journal of economics*, 114, 1243–1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic growth*, 8, 155–194.

Asher, S., T. Lunt, R. Matsuura, and P. Novosad (2021a): "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform," *The World Bank Economic Review*, 35, 845–871.

——— (2021b): "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform," *The World Bank Economic Review*, 35.

Banerjee, A. and R. Somanathan (2007): "The political economy of public goods: Some evidence from India," *Journal of development Economics*, 82, 287–314.

Baskaran, T., S. Bhalotra, B. Min, and Y. Uppal (2023): "Women legislators and economic performance," *Journal of Economic Growth*, 1–64.

Bharathi, N., D. Malghan, and A. Rahman (2023): "Ethnic diversity and economic development with spatial segregation," *Economics Letters*, 222, 110951.

Bove, V. and L. Elia (2017): "Migration, diversity, and economic growth," *World Development*, 89, 227–239.

Bredtmann, J., K. Nowotny, and S. Otten (2020): "Linguistic distance, networks and migrants' regional location choice," *Labour Economics*, 65, 101863.

Chiswick, B. R. and P. W. Miller (2005): "Linguistic distance: A quantitative measure of the distance between English and other languages," *Journal of multilingual and multicultural development*, 26, 1–11.

Desmet, K., I. Ortuño Ortín, and S. Weber (2005): "Peripheral diversity and redistribution," .

Desmet, K., S. Weber, and I. Ortuño-Ortín (2009): "Linguistic diversity and redistribution," *Journal of the European Economic Association*, 7, 1291–1318.

Duranton, G. and D. Puga (2004): "Micro-foundations of urban agglomeration economies," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2063–2117.

Dyen, I., J. B. Kruskal, and P. Black (1992): "An Indoeuropean classification: A lexicostatistical experiment," *Transactions of the American Philosophical society*, 82, iii–132.

Easterly, W. and R. Levine (1997): "Africa's growth tragedy: policies and ethnic divisions," *The quarterly journal of economics*, 1203–1250.

Elvidge, C. D., P. Cinzano, D. Pettit, J. Arvesen, P. Sutton, C. Small, R. Nemani, T. Longcore, C. Rich, J. Safran, et al. (2007): "The Nightsat mission concept," *International Journal of Remote Sensing*, 28, 2645–2670.

Farr, T. G. and M. Kobrick (2000): "Shuttle Radar Topography Mission produces a wealth of data," *Eos, Transactions American Geophysical Union*, 81, 583–585.

Fearon, J. D. (2003): "Ethnic and cultural diversity by country," *Journal of economic growth*, 8, 195–222.

Fearon, J. D. and D. D. Laitin (2003): "Ethnicity, insurgency, and civil war," *American political science review*, 97, 75–90.

Ghani, E., W. Kerr, and S. D. O'Connell (2014): "What makes cities more competitive? Lessons from India," *Economic Premise*, 132, 1–4.

Greenberg, J. H. (1956): "The measurement of linguistic diversity," *Language*, 32, 109–115.

Henderson, J. V., A. Storeygard, and D. N. Weil (2011): "A Bright Idea for Mesuring Economic Growth," *American Economic Review*.

——— (2012): "Measuring economic growth from outer space," *American economic review*, 102, 994–1028.

Henderson, V. (2003): "The urbanization process and economic growth: The so-what question," *Journal of Economic growth*, 8, 47–71.

Jain, T. (2017): "Common tongue: The impact of language on educational outcomes," *The Journal of Economic History*, 77, 473–510.

Krugman, P. (1991): "Increasing returns and economic geography," *Journal of political economy*, 99, 483–499.

Montalvo, J. G. and M. Reynal-Querol (2005): "Ethnic diversity and economic development," *Journal of Development economics*, 76, 293–323.

——— (2021): "Ethnic diversity and growth: Revisiting the evidence," *Review of Economics and Statistics*, 103, 521–532.

Ottaviano, G. I. and G. Peri (2005): "Cities and cultures," *Journal of Urban Economics*, 58, 304–337.

Ratna, N. N., R. Q. Grafton, and T. Kompas (2009): "Is diversity bad for economic growth?: Evidence from state-level data in the US," *The Journal of Socio-Economics*, 38, 859–870.

Redding, S. J. and E. Rossi-Hansberg (2017): "Quantitative spatial economics," *Annual Review of Economics*, 9, 21–58.

Shannon, C. E. (1948): "A mathematical theory of communication," *The Bell system technical journal*, 27, 379–423.

Family	Number of Speakers	Percentage
Indo-European	945,333,910	78.07
Dravidian	237,840,116	19.64
Austro-Asiatic	13,493,080	1.11
Sino-Tibetan	12,257,382	1.01
Afro-Asiatic	54,947	0.00

Table 1: Language Families of India

Table 2: Summary Statistics

	2001			2011		
Variable	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Log Per Capita Nighttime Lights	−13.06	−13.11	1.19	−12.39	−12.56	1.04
Greenberg Index	0.29	0.24	0.24	0.26	0.19	0.23
B-Index	0.24	0.18	0.20	0.59	0.25	0.13
Sex Ratio	0.48	0.48	0.02	0.49	0.49	0.02
SC Population Ratio	0.14	0.15	0.09	0.15	0.16	0.09
Literacy Rate	0.54	0.53	0.12	0.63	0.62	0.11
Terrain Ruggedness Index	10.05	5.28	10.90	10.05	5.28	10.90
Distance to Coast (km)	496.83	412.03	385.81	496.83	412.03	385.81
Observations		N			N	

Notes: This table presents summary statistics for key variables used in the analysis for 2001 and 2011. Log Per Capita Nighttime Lights refers to the natural logarithm of nighttime light intensity per capita. The Greenberg Index measures linguistic diversity. B-Index represents the diversity index that incorporates linguistic distance between languages. Sex Ratio is defined as the proportion of females in the population. SC Population Ratio is the proportion of Scheduled Caste population. Literacy Rate is the proportion of literate population aged 7 and above. Terrain Ruggedness Index measures topographical variation. Distance to Coast is measured in kilometers from the nearest coastline.

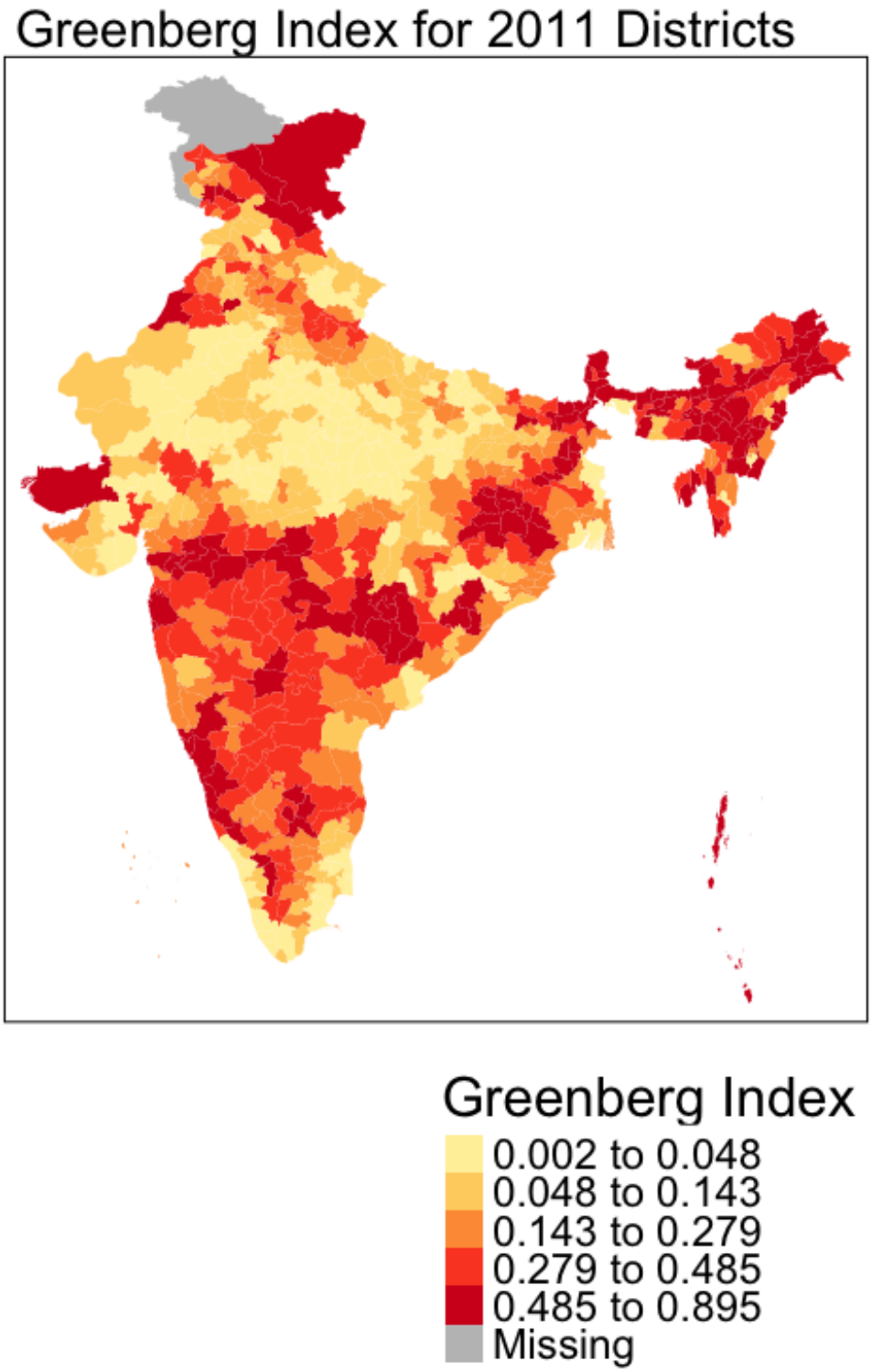


Figure 1: Greenberg Index for the districts of India, 2011

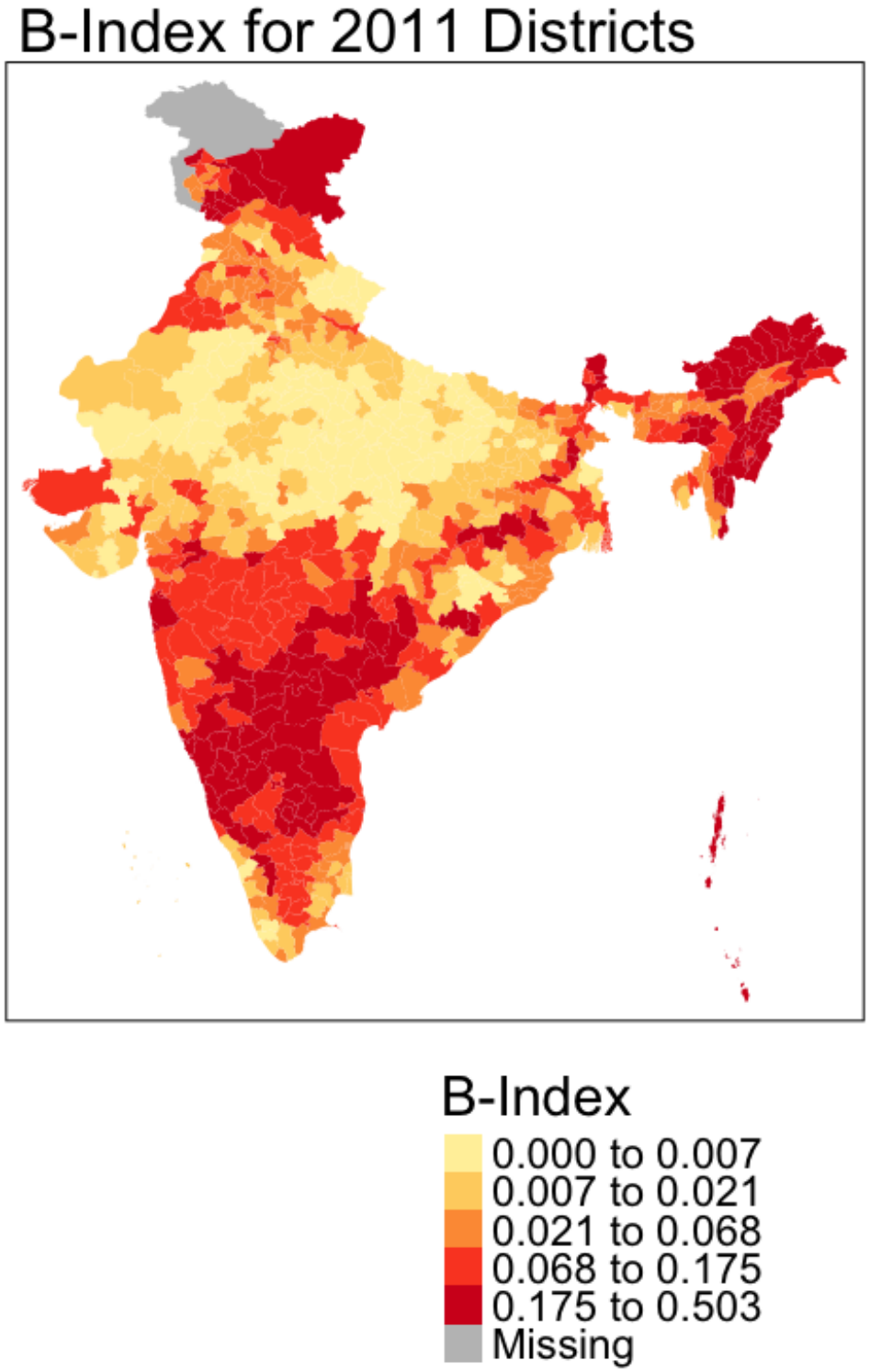


Figure 2: B-Index for the districts of India, 2011

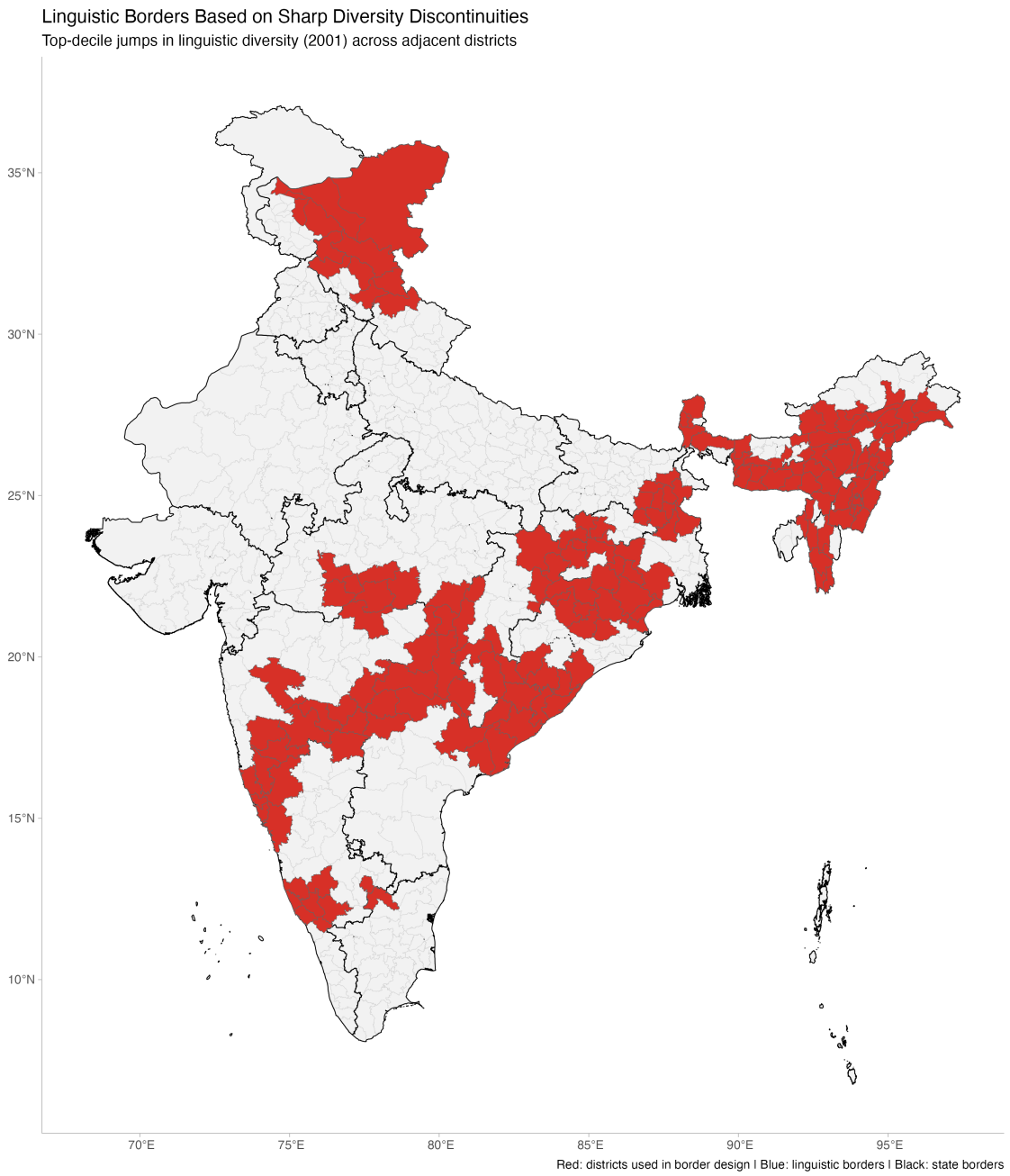


Figure 3: Border-pair design based on linguistic diversity
The figure illustrates the border-pair research design. The map shows adjacent administrative districts. Border-pairs are constructed from districts that share a common administrative boundary. The empirical analysis restricts attention to border-pairs in the top 10 percent of the distribution of the absolute difference in the linguistic B-Index across all district border-pairs, ensuring large discontinuities in linguistic composition across neighboring districts.

Table 3: Greenberg Index and Growth in Per Capita Nighttime Lights

Dependent Variable:	d_log_pcnl			
	Model 1	Model 2	Model 3	Model 4
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	0.48***	2.00	0.87	
	(0.11)	(1.83)	(1.56)	
Greenberg Index	-0.43	-0.95	-0.55	-0.89***
	(0.68)	(0.58)	(0.36)	(0.31)
Greenberg Index^2	2.24**	2.16**	1.35***	1.35***
	(0.92)	(0.81)	(0.48)	(0.36)
PC NTL		-1796.65*	-360.73	-4.13
		(888.76)	(676.72)	(773.81)
Sex Ratio		0.82	1.61	-0.42
		(3.47)	(2.94)	(1.93)
SC Ratio		-3.37***	-1.66**	-0.92
		(1.08)	(0.65)	(0.76)
Literacy Rate		-2.35***	-2.55***	-1.97***
		(0.51)	(0.25)	(0.32)
Terrain Ruggedness			0.04***	0.04***
			(0.00)	(0.01)
Log Distance to Coast			0.03	0.06
			(0.05)	(0.04)
Fixed-effects				
State				Yes
Fit statistics				
R^2	0.13	0.33	0.57	0.69
Adjusted R^2	0.13	0.33	0.56	0.67

Clustered (State) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Clustered standard errors at the state level (pc11_state_id) are shown in parentheses.
Significance codes: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: B-Index and Growth in Per Capita Nighttime Lights

Outcome:	d_log_pcnl				
	Model 1	Model 2	Model 3	Model 4	Model 5
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Constant	0.47***	1.94	0.90		
	(0.11)	(1.81)	(1.51)		
B-Index	-0.47	-1.09	-0.56	-0.79**	-0.84**
	(0.79)	(0.67)	(0.42)	(0.38)	(0.40)
B-Index^2	3.22***	3.07***	1.78***	1.53***	1.72***
	(1.08)	(1.00)	(0.59)	(0.46)	(0.49)
PC NTL		-1,582.10*	-257.20	-2.11	-2,403.70
		(909.10)	(686.90)	(763.60)	(2,655.40)
Sex Ratio		0.82	1.51	-0.55	-0.36
		(3.39)	(2.85)	(1.91)	(2.57)
SC Ratio		-3.24***	-1.64**	-0.87	-1.08
		(1.10)	(0.66)	(0.75)	(0.78)
Literacy Rate		-2.29***	-2.51***	-1.95***	-1.65***
		(0.49)	(0.24)	(0.31)	(0.37)
Terrain Ruggedness			0.04***	0.04***	0.04***
			(0.00)	(0.01)	(0.01)
Log Distance to Coast			0.03	0.05	0.09
			(0.05)	(0.04)	(0.06)
Log Distance to Capital (km)					-0.05
					(0.05)
Fixed-effects					
State				Yes	Yes
Fit statistics					
R^2	0.16	0.35	0.57	0.69	0.70
Adjusted R^2	0.16	0.34	0.57	0.67	0.68

Clustered (State) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Notes: Clustered standard errors at the state level (pc11_state_id) are shown in parentheses.
Significance codes: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Shannon Index and Growth in Per Capita Nighttime Lights

Dependent Variable:	d_log_pcnl			
	Model 1	Model 2	Model 3	Model 4
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	0.47***	1.95	0.89	
	(0.11)	(1.82)	(1.51)	
Shannon Index	-0.43	-0.94	-0.55	-0.88***
	(0.68)	(0.58)	(0.36)	(0.31)
Shannon Index^2 (2001)	2.24**	2.15**	1.35***	1.35***
	(0.91)	(0.81)	(0.48)	(0.36)
Per Capita NTL		-1,780*	-360	-4.10
		(885)	(675)	(775)
Sex Ratio		0.82	1.61	-0.42
		(3.47)	(2.95)	(1.93)
SC Ratio		-3.36***	-1.66**	-0.92
		(1.08)	(0.65)	(0.76)
Literacy Rate		-2.35***	-2.55***	-1.97***
		(0.51)	(0.25)	(0.32)
Terrain Ruggedness			0.04***	0.04***
			(0.00)	(0.01)
Log Distance to Coast			0.03	0.06
			(0.05)	(0.04)
Fixed-effects				
State				Yes
Fit statistics				
R^2	0.13	0.33	0.57	0.67
Adjusted R^2	0.13	0.33	0.56	0.67

Clustered (State) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Clustered standard errors at the state level (pc11_state_id) are shown in parentheses.
Significance codes: *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Trimmed Models

Dependent Variable:	d_log_pcnl	
	Top 5% Removed	Top 10% Removed
Variables		
Constant	0.85***	0.83***
	(0.31)	(0.33)
B-Index	-0.78**	-0.76**
	(0.37)	(0.38)
B-Index squared	1.52***	1.49***
	(0.46)	(0.48)
PC NTL	-3.98	-4.12
	(0.77)	(0.79)
Sex Ratio	-0.42	-0.39
	(1.92)	(2.01)
SC Ratio	-0.92	-0.91
	(0.76)	(0.78)
Literacy Rate	-1.97***	-1.96***
	(0.32)	(0.33)
Terrain Ruggedness	0.0409***	0.0398***
	(0.0061)	(0.0063)
Log Distance to Coast	0.0646	0.0621
	(0.0432)	(0.0445)
Fixed-effects		
pc11_state_id	Yes	Yes
Fit statistics		
R^2	0.69	0.69
Adjusted R^2	0.67	0.66

Clustered (pc11_state_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Clustered standard errors at the state level (pc11_state_id) are shown in parentheses.
Significance codes: *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Linguistic diversity and local growth: Border-pair estimates

	(1)	(2)	(3)
Linguistic Diversity (B-Index, 2001)	0.014***	0.012***	0.011**
	(0.005)	(0.004)	(0.005)
Initial Nighttime Lights (log)	-0.020***	-0.018***	-0.017***
	(0.005)	(0.004)	(0.004)
Geographic Controls	No	Yes	Yes
Border-pair Fixed Effects	Yes	Yes	Yes
Bandwidth (km)	50	50	25
Observations	612	612	328
Border-pairs	306	306	164

[flushleft]

Notes: The dependent variable is the change in log nighttime lights per capita between 2001 and 2011. The sample includes only administrative district border-pairs in the top 10 percent of the distribution of the absolute difference in the linguistic B-Index across all border-pairs. All specifications include border-pair fixed effects, comparing adjacent districts on opposite sides of the same administrative boundary. Standard errors are clustered at the border level.

*** p<0.01, ** p<0.05, * p<0.1.

Table 8: Linguistic Diversity, Market Thickness, and Firm Structure

	Association with Linguistic Diversity	
Outcome	Direction	Interpretation
Market thickness		
Non-farm employment density (2013)	+	Denser local labour markets
Establishment growth (2005–2013)	+	Greater entry and expansion
Firm structure		
Share of small establishments (<20 workers)	−	Reduced prevalence of micro firms
Share of large establishments (≥100 workers)	+	Greater viability of scale

Notes: The table summarises patterns from district-level regressions using data from the Economic Census of India (2005, 2013). All specifications control for initial development, demographic characteristics, and state fixed effects. Positive (+) and negative (−) signs indicate statistically significant associations.