Completing the Incomplete:

Socioeconomic Ranking through Unsupervised Clustering

Ayushman Roy

CS-2378/ POL-2070/ MS-3511-1 The New Geography of the Information Age

Instructor: Professor Debayan Gupta

1. Introduction

The Human Development Index (HDI), the Multidimensional Poverty Index (MPI), and most other socioeconomic indices use a weighted average of socioeconomic indicators for assigning scores. This methodology is inherently normative. This methodology implicitly assumes that socioeconomic improvement relies solely on the selected indicators. Consequently, the validation of the resultant scores then becomes a question regarding the applicability and validity of selected indicators and weights for deciding socioeconomic status. This methodology imposes an inherent bias on such scores and rankings, and consequently, the ranks individual datapoints get is an evaluation of their current attempt and trajectory in chasing the constructed idealistic goal. These current methodologies ignore the empirical differences present within the underlying data for assigning the scores and the ranks.

I attempt to make a socioeconomic index that bypasses such selection of socioeconomic indicators. I intend to let the underlying empirics drive the segregation and consequent scoring of individual datapoints. I exploit relative differences between datapoints through unsupervised machine learning methods (*k-means clustering*) for scoring and ranking individual districts. I further analyze the resultant socioeconomic scores and rankings to evaluate whether the biases inherent in conventional methodologies introduce significant distortions in the representation of actual socioeconomic hierarchies. This evaluation is undertaken through a comparative analysis of the rankings generated by the paper's unsupervised machine learning approach, the HDI rankings, and other established indicators of well-being. We gather insights into whether conventional methods, which rely heavily on predefined indicators and normative weightings, accurately capture the underlying socioeconomic realities or whether they impose artificial structures on the data.

A critical aspect of our evaluation involves examining the implications of relying exclusively on the deviation between an idealized benchmark and the observed characteristics of individual districts, a central assumption in conventional index construction. I investigate whether an alternative assumption, that districts with similar socioeconomic profiles naturally form distinct groups and the identity of these distinct groups better capture the structure of socioeconomic variation. Additionally, we examine the relative efficacy of two empirical approaches to index construction: (i) the use of Principal Component Analysis (PCA) for both

variable selection and weighting based on the underlying variance structure, and (ii) the use of the full high-dimensional dataset without dimensionality reduction, allowing all observed socioeconomic indicators to contribute to the clustering and scoring process.

We evaluate these approaches against conventional normative methods for assessing whether empirically grounded techniques provide a more accurate and unbiased representation of socioeconomic status, or whether they introduce new limitations and uncertainties into the measurement process.

2. Literature Review

The Multidimensional Poverty Index (MPI) has been developed to provide a comprehensive measure of poverty beyond income-based metrics. Introduced by the United Nations Development Programme (UNDP), the MPI captures deprivation across three fundamental dimensions, health, education, and standard of living, using a set of ten indicators. These indicators range from child mortality and malnutrition to access to sanitation, electricity, and clean cooking fuel. The MPI adopts the Alkire-Foster methodology, which first identifies households as deprived across individual indicators and then classifies households as multidimensionally poor if they experience deprivation in one-third or more of the weighted indicators (UNDP, 2021). The MPI is computed as the product of the incidence of poverty (H), the proportion of the population identified as multidimensionally poor, and the intensity of poverty (A), which reflects the average proportion of deprivations experienced by the poor. By incorporating intensity, the MPI captures the depth of poverty, ensuring that policies do not merely focus on borderline cases but also target households experiencing acute deprivation. The index has remained structurally stable since 2018, maintaining equal weights across its three dimensions while applying discrete cutoffs to classify populations into vulnerable, poor, and severely poor categories (UNDP, 2021).

The MPI exhibits several methodological limitations that stem primarily from its reliance on discrete categorical thresholds and predefined indicator selection and weighting schemes. These normative design choices constrain the MPI's flexibility in capturing the full spectrum of socioeconomic conditions, especially in contexts where deprivation manifests in complex or locally specific ways. Furthermore, by assigning equal weights to dimensions and fixed cutoffs to classify poverty, the MPI may obscure important relative differences between households or

regions. This approach prioritizes conformity to an externally imposed ideal over the empirical diversity present within the underlying data.

In parallel, the Human Development Index (HDI) offers a broader measure of human development, incorporating life expectancy, education, and income per capita into a composite score. The HDI employs goalposts, or minimum and maximum values, to normalize indicators and produce standardized dimension indices. The final HDI score is computed as the geometric mean of these indices, reflecting a more balanced treatment of dimensions compared to simple averages. However, like the MPI, the HDI relies heavily on normative assumptions about what constitutes desirable development and how progress should be measured across diverse contexts (UNDP, 2021-22). Extensions of the HDI, such as the Inequality-adjusted HDI (IHDI), attempt to account for inequalities within each dimension by discounting average achievements based on observed inequality levels. This adjustment introduces an important distributional perspective, ensuring that societies with higher internal disparities receive lower development scores, even if their average performance remains high. Similarly, the Planetary Pressures-adjusted HDI (PHDI) extends this logic to account for environmental sustainability, discounting human development achievements based on ecological impacts and resource consumption. Both adjustments reflect growing recognition of the need to incorporate equity and sustainability considerations into development assessments, yet they retain the core structure and normative indicator set of the original HDI (UNDP, 2021-22).

3. Methodology

data_processor.py → {geo_mapper.py, ideal_district.py} → analysis.py → ranking.py → {ranking_mapper.py, ranking_analysis.py} → {validation_mapper.py}

3.1. Data Processing

We primarily leverage Periodic Labor Force Survey (PLFS) and National Family Health Survey (NFHS) datasets for constructing the index construction dataset. Specifically, we use NFHS 5 aggregated at the district-level, and hhv1 and perv1 datasets from PLFS. The project also generates and utilizes a distance variable that calculates the district's distance from tier-1 cities of India for index construction. Additionally, we also use district-level HDI data from Indian Journal of Human Development, district-level road statistics data from Ministry of Road

Transport and Highways of India, and district-level projected population growth data from International Institute for Population Sciences for ranking validation purposes.

Firstly, we clean variables of NFHS 5 manually. Some variables like *Unmet need for spacing, Institutional births in private facility (in the 5 years before the survey) (%),* etc. were removed because they were already incorporated within other socioeconomic indicators like *unmet family planning needs*, *public institutional births*, and *total institutional births.* Some variables like *Children born at home who were taken to a health facility, Births in a private health facility that were delivered by caesarean section,* etc. were removed because of many missing values. Some variables which included specific vaccination figures, segregation between diets of breastfeeding and non-breastfeeding children, etc. were removed because of overfitting concerns and possible multicollinearity. Additionally, the following variables were dropped based on Spearman correlation with a 0.9 threshold: *Women (age 15-49) who are literate (%), Mothers who received postnatal care from a doctor/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (for last birth in the 5 years before the survey) (%), Institutional births (in the 5 years before the survey) (%), Female population age 6 years and above who ever attended school (%), Births attended by skilled health personnel (in the 5 years before the survey) (%), Children who received postnatal care from a doctor / nurse / LHV / ANM / midwife / other health personnel within 2 days of delivery (for last birth in the 5 years before the survey)*. Additionally, NSS (National Sample Survey) region codes are imported and merged with the cleaned dataset based on state and district identifiers. Subsequently, the PLFS datasets are cleaned.

For the household-level dataset (hhv1), a weight variable is calculated by dividing the multiplier (*mult*) by the number of quarters (*no_qtr*) to account for annual estimates. District-level aggregates are then computed using custom aggregation functions. These include weighted means and medians for continuous variables such as household size (*hh_size*) and various consumption expenditure categories (*hce1, hce2*, etc.), as well as proportional distributions for categorical variables like religion (*relg*) and social group (*sg*). Categorical distributions are expanded into separate columns representing percentage shares for each category (e.g., religion or social group proportions). For the person-level dataset (perv1), additional preprocessing is performed, including the creation of age groups using predefined bins (e.g., <18, 18-35, etc.).

Weighted means and medians are calculated for continuous variables such as education levels (*form_edu*) and earnings (*ern_reg, ern_self*). For categorical variables (e.g., marital status, educational levels, vocational training details, employment types), weighted distributions are computed to capture the proportional representation of each category within districts.

Additionally, OpenCage is used for generating the distance variable which calculates the distance of each district from the nearest tier-1 city (Ahmedabad, Bengaluru, Chennai, Delhi, Hyderabad, Kolkata, Mumbai, and Pune). Once district-level aggregates are generated for both household- and person-level datasets, they undergo a cleaning process to ensure data quality. Columns with insufficient coverage (valid data ≤ 67% of districts) are dropped from each dataset, ensuring only reliable and representative variables are retained for further analysis. The cleaned household-level (PLFS_hhv1.csv), person-level (PLFS_perv1.csv), and enriched NFHS master datasets are merged into a unified master file based on common state and district identifiers.

3.2. Data Analysis and Index Construction

The master sheet underwent comprehensive preprocessing to standardize variable formats and handle missing data. Non-numeric entries, including placeholders such as asterisks and blanks, were replaced with NaN values, and descriptive identifier columns (state and district names, codes, and pre-existing HDI values) were isolated from the analytical variables to ensure analytical independence. The first major step involved removing highly collinear variables to minimize redundancy and multicollinearity bias. A Spearman correlation matrix was computed to capture pairwise monotonic relationships between all variables, and any variable with an absolute correlation coefficient exceeding 0.90 with another variable was flagged for removal. This variable selection process ensured that the retained indicators contributed unique information to the subsequent analysis. The remaining variables were standardized using Z-score normalization, transforming each variable to have a mean of zero and a standard deviation of one. This step ensured comparability across indicators with different scales, units, and ranges. Missing values were then imputed using a K-Nearest Neighbors (KNN) imputation method, which operates on the assumption that districts with similar characteristics across observed variables will exhibit similar values for missing variables. The imputation process weighted the contributions of neighboring observations by their proximity, ensuring that imputations reflected

socioeconomic similarity. Then we generated the Pearson correlation matrix (post-imputation and normalization), and descriptive statistics for these correlations, including mean, median, and proportions exceeding certain thresholds (e.g., 0.50, 0.75, 0.90), were computed for assessing overall variable interrelatedness.

PCA was applied in two configurations: (i) a simplified three-component PCA for exploratory analysis and interpretation, and (ii) an optimal-component PCA capturing ≥80% of the total explained variance. The optimal number of components (n=43) was determined via a scree plot, which mapped the cumulative explained variance against the number of components, with an 80% threshold serving as the cutoff for dimensionality reduction. To facilitate relative comparisons, all variables, including the raw (vanilla) indicators, the three-component PCA scores, and the optimal-component PCA scores, got rescaled using Min-Max normalization, ensuring that all dimensions ranged between 0 and 1. The empirical methodology relied on the construction of a hypothetical *ideal* district, a synthetic benchmark entity representing optimal performance across all normalized dimensions (using the logic specified in LEGEND.xlsx). For each representation (vanilla, PCA-3, PCA-optimal), Euclidean distances were computed between each district and the ideal district, producing a set of distance scores that quantified the relative developmental gap between each real-world district and the aspirational ideal.

These transformed datasets were then subjected to K-means clustering which grouped districts into clusters based on their multidimensional proximity. To ensure robustness, clustering was performed across three train-test split configurations (33%-67%, 67%-33%, and 100% train), allowing assessment of cluster stability across varying data partitions. Both the Elbow Method and Silhouette Scores were used to guide the selection of the optimal number of clusters (k=7). Additionally, for each selected cluster configuration (k = 2, 4, 7, 10), the clustering results were combined with district identifiers, and the Euclidean distance between each district's cluster centroid and the ideal district was computed, yielding an additional measure of cluster-level developmental proximity. Cluster-level Human Development Index (HDI) scores were also computed to explore the relationship between cluster membership and developmental outcomes.

3.3. Ranking

The core of the ranking methodology relies on leveraging cluster-level for contextualizing each district’s development standing relative to its assigned cluster. For each

clustering configuration, districts are first assigned a cluster-level Human Development Index (HDI) score. This cluster HDI is calculated as the average HDI of all districts belonging to the same cluster. Assigning this average HDI serves two purposes: it reflects the general developmental context of the cluster, and it allows for comparisons between districts within the same clustering configuration by incorporating the shared developmental environment into their rankings. Subsequently, a two-stage normalization process (Z-score, Min-Max) is applied to a series of cluster-level metrics, specifically including centroid distances, within-cluster squared distances, and cluster-level HDI.

An additional transformation step is applied to capture each district's proximity to the ideal district within each PCA configuration. For each district under each clustering configuration, the Euclidean distance to this ideal district is computed, with smaller distances indicating greater proximity to the ideal developmental profile. To ensure these distances could be meaningfully incorporated into composite scores, they are transformed using the formula $x' = \frac{1}{1+x}$, where greater proximity (i.e., lower distance) maps to higher values on a normalized [0, 1] scale. In addition to proximity to the ideal district, cluster stability is explicitly incorporated into the ranking process through the calculation of Within-Cluster Squared Distances (WCSD). The WCSD is a part-stability metric that captures the sum of squared distances between each cluster's centroid and the ideal district vector for each district under each PCA configuration. Clusters with higher WCSD are considered more stable because (i) their centroids lie closer to the optimal developmental benchmark reflecting tighter alignment with desirable socioeconomic conditions, and (ii) the number of districts in the cluster is high.

$$cluster\ scores = x_1 \times centroid\_dist + x_2 \times centroid\_dist^2 \times cluster\_count$$

To derive configuration-specific composite performance scores for each district, three metrics are combined: (i) WCSD (indicating cluster stability), (ii) cluster centroid's transformed proximity from the ideal district (indicating score against optimal developmental benchmark), and the cluster's average HDI (capturing the general developmental context of the cluster). These components are weighted at 0.3, 0.3, and 0.4 respectively, with greater emphasis placed on the cluster HDI to prioritize contextual development conditions while retaining sensitivity to cluster cohesion and ideal proximity. These configuration-specific composite scores are calculated for each district across all clustering configurations. Finally, a blending step is performed to

integrate individual district scores across configurations. The cluster-based composite score for each district is merged with the previously computed individual proximity from ideal district through weights *alpha* and *beta*. This blending ensures that the final socioeconomic score for each district reflects both (i) positioning within the broader cluster structure and (ii) direct alignment with the optimal district profile. This blended score is then used for generating district-level socioeconomic rankings for each configuration, higher scores indicating more favorable socioeconomic conditions. To assess the stability and internal consistency of these rankings across configurations, a Spearman rank correlation matrix was computed. This final step serves as a robustness check, assessing confidence on the reliability and generalizability of the derived district-level rankings across varying clustering scenarios.

4. Analysis of Constructed Rankings

alpha = 0.5, beta = 0.5	
Top Ranks	**Worst Ranks**
3_socioeconomic_scores	29_socioeconomic_ranks
6_socioeconomic_scores	32_socioeconomic_ranks
12_socioeconomic_scores	35_socioeconomic_ranks
9_socioeconomic_scores	27_socioeconomic_ranks
24_socioeconomic_scores	34_socioeconomic_ranks

alpha = 0.25, beta = 0.75	
Top Ranks	**Worst Ranks**
3_socioeconomic_scores	32_socioeconomic_ranks
6_socioeconomic_scores	35_socioeconomic_ranks
12_socioeconomic_scores	27_socioeconomic_ranks
9_socioeconomic_scores	29_socioeconomic_ranks
18_socioeconomic_ranks	31_socioeconomic_ranks

alpha = 0, beta = 1	
Top Ranks	**Worst Ranks**
1_socioeconomic_scores	25_socioeconomic_ranks
2_socioeconomic_scores	26_socioeconomic_ranks
3_socioeconomic_scores	27_socioeconomic_ranks
4_socioeconomic_scores	28_socioeconomic_ranks
5_socioeconomic_scores	29_socioeconomic_ranks

alpha = 0.9, beta = 0.1	
Top Ranks	**Worst Ranks**
3_socioeconomic_ranks	32_socioeconomic_ranks
6_socioeconomic_ranks	34_socioeconomic_ranks
24_socioeconomic_ranks	35_socioeconomic_ranks
12_socioeconomic_ranks	23_socioeconomic_ranks
9_socioeconomic_ranks	31_socioeconomic_ranks

The analysis of constructed rankings focuses on evaluating the performance of 36 distinct ranking configurations. The objective is to systematically test how these methodological choices impact the quality of the final district-level rankings. To assess ranking quality, the correlation between each configuration's district rankings and HDI scores was used as a proxy. The underlying logic is that a robust and theoretically sound socioeconomic index should exhibit a reasonably strong positive correlation with HDI.

Ranking configurations 3, 6, 9, and 12 consistently demonstrated the highest positive correlations with HDI, indicating that these approaches produced rankings most aligned with established measures of human development. These configurations share two features: (i) they are based on the vanilla dataset, and (ii) they use the complete dataset for k-means clustering (100%-0% train-test split). They differ only in the number of clusters (k=2, 4, 7, 10). This highlights the robustness of using the full dataset without PCA transformation and suggests that this basic configuration offers a strong foundation for constructing district-level socioeconomic rankings. Among these, Ranking 12 (with k=10) was selected as the primary configuration for subsequent validation and analysis due to its strong HDI correlation and stable performance across sensitivity checks. Configuration 24 also performed well, although it represents an outlier because it uses PCA with three components.

A particularly striking finding relates to the overall underperformance of rankings derived from PCA-transformed datasets, especially those using a large number of principal components. In several cases, these PCA-based configurations exhibited negative correlations with HDI, indicating that the derived socioeconomic rankings were not only misaligned with HDI, but in some cases inversely related. This severe underperformance can be explained by the inherent limitations of applying dimensionality reduction techniques to datasets designed to capture complex socioeconomic phenomena. Socioeconomic indices typically rely on theoretical

frameworks to guide variable selection, ensuring that key indicators representing dimensions such as health, education, employment, and infrastructure are appropriately represented. In contrast, PCA is purely empirical, identifying axes of maximum variance without regard to substantive meaning. This process can overweight variables with high variance that may have limited conceptual relevance (e.g., specific lifestyle behaviors or localized economic conditions), while underweighting variables that are theoretically essential (e.g., gender ratios, infant mortality rates, or educational attainment). Consequently, the principal components selected through this process may inadequately capture core elements of socioeconomic development, severely compromising the interpretability and validity of the final rankings.

PCA_3 Top Components: *sg_1_pct, marst_1, fam_female_ster, women_tobbaco, men_alc, delivery_expense, men_tobbaco, form_edu_mean, wrkr_sas_1.0, hh_size_mean, marst_3, ssec_pas_7.0, mother_folic, gedu_lvl_12, child_anaemic*

PCA_Optimal Top Components: *sex_ratio_birth, wrk_365_1.0, age_group_age_18-35, gedu_lvl_6, voc_6.0, eff_pas_1.0, hh_salt, etyp_sas_1.0, women_oral, missing_count, etyp_sas_2.0, child_resp, child_overwght, acws_95, hh_drnk_water*

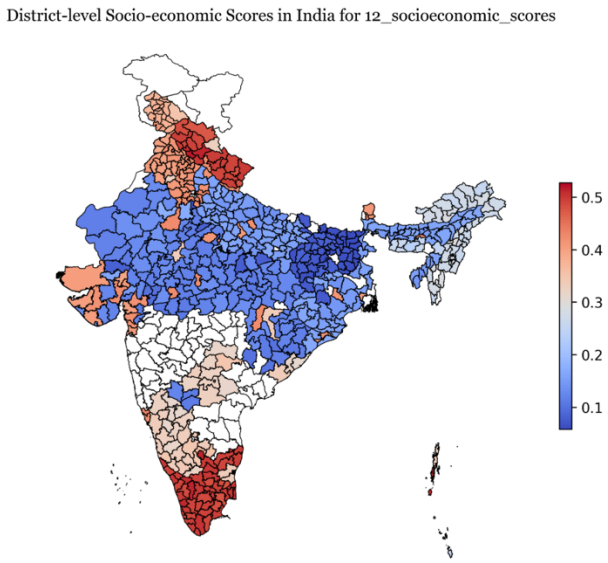
An additional layer of analysis focused on the top-ranked variables contributing to the principal components in the PCA-based configurations. This comparison revealed no overlap between the top variables in the PCA-3 configuration and the PCA-optimal configuration. The complete lack of overlap between these two lists highlights a core limitation of unsupervised dimensionality reduction: the identified components are highly contingent on the specific method of selection and the criteria applied, making them unstable across configurations and difficult to interpret in any coherent theoretical framework. This instability further erodes the substantive interpretability of PCA-based rankings, making them unsuitable for contexts where theoretical validity and conceptual clarity are paramount.

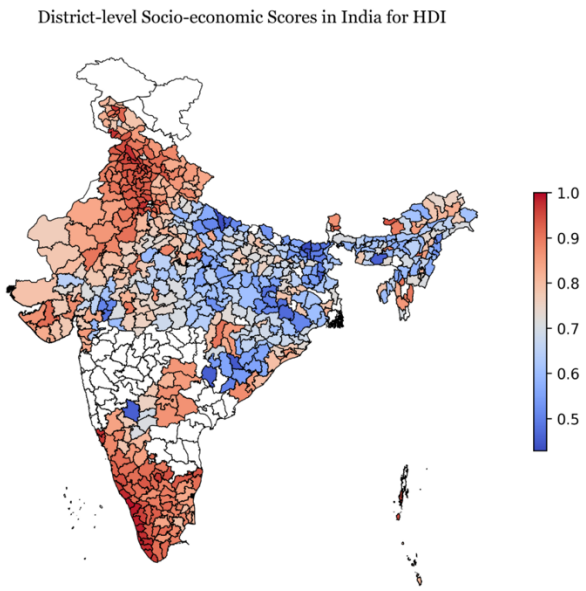
The impact of splitting the dataset into train and test subsets for clustering was also found to be negative, though less severe than the effect of PCA. Configurations that employed train-test splits (33%-67% or 67%-33%) systematically underperformed those using the full dataset (100%-0%). This suggests that excluding a portion of districts from the clustering process reduces the ability of k-means to fully capture the underlying structure of socioeconomic variation across all districts. Clustering on a subset necessarily omits potentially important

district-level patterns, resulting in clusters that are less representative and stable. This effect is particularly concerning in the context of district-level analysis, where the number of observations is inherently limited, and each district represents a unique socio-economic profile that contributes to the broader structure of variation. The loss of this diversity weakens cluster coherence, which in turn undermines the reliability of the final rankings.

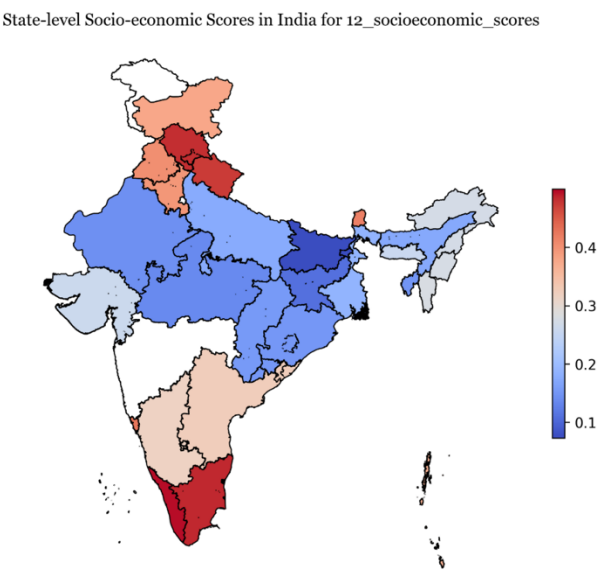
The effect of varying the number of clusters in the ranking configurations was more ambiguous. A comparison of HDI correlations across Configurations 3, 6, 9, and 12 (all sharing the same vanilla data treatment and 100-0 train-test split) but vary by cluster count (k=2, 4, 7, 10) reveals a slight, though non-linear, improvement in HDI correlation as the number of clusters increases. Specifically, the HDI correlation rises from 0.66 at k=2 (Configuration 3) to 0.72 at k=7 (Configuration 9), before stabilizing. This suggests a modest benefit to increasing the granularity of clusters, potentially reflecting greater flexibility in capturing localized socioeconomic variation within smaller, more homogenous groupings. However, the absence of a clear linear trend implies that this benefit plateaus after a certain point, likely because excessive fragmentation reduces the interpretability of clusters and introduces noise into the final ranking process.

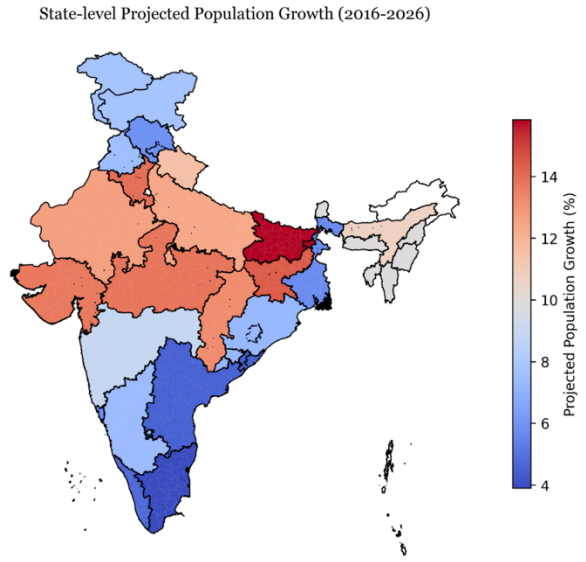
In conclusion, the analysis highlights three key insights: (i) PCA-based transformations introduce substantial risks to ranking validity due to loss of meaningful variation and theoretical coherence; (ii) train-test splits impair the stability of clustering and ranking processes, and (iii) increasing cluster counts offers limited but non-negligible benefits to ranking reliability, provided theoretical consistency is maintained.





This visualization of Ranking Configuration #12 alongside the Human Development Index (HDI) reveals a strong positive correlation and partially validate the rankings, demonstrating that higher-ranked states tend to align with higher HDI values.





States with high projected population growth show a notable negative correlation with the rankings. This trend reaffirms the established inverse relationship between socioeconomic status and fertility rates, where regions with lower socioeconomic development often experience higher population growth.

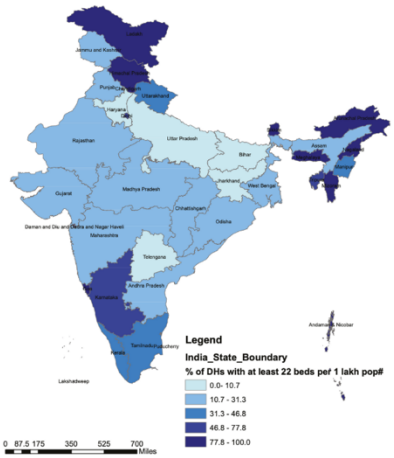
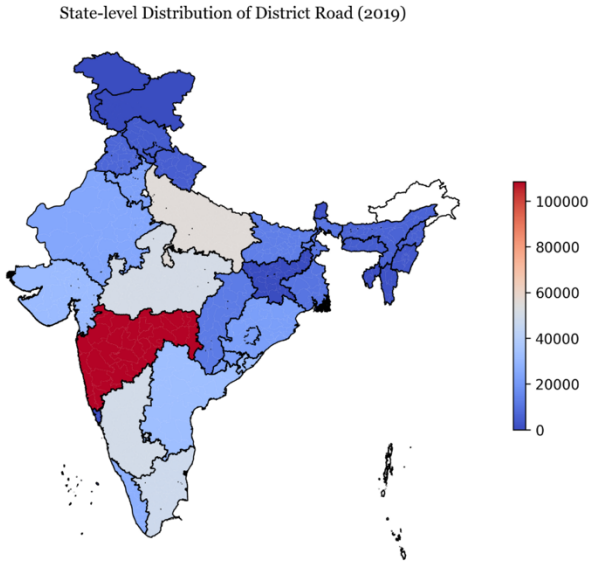


Figure 3.1.2: Percentage of district hospitals in each State/UT that have at least 22 beds per 1 lakh population



There is also a significant correlation between the rankings and district-level hospital bed availability, suggesting that healthcare infrastructure plays a critical role in socioeconomic status. However, road distance per state does not appear to align with the rankings. This discrepancy may arise from road distance being influenced by government budgets, an external factor that might not accurately reflect socioeconomic conditions.

5. Limitations and Areas of Future Research

While the methodology employed in this study offers a flexible framework for constructing district-level socioeconomic rankings, several limitations emerge that point to areas requiring further research and methodological refinement. One key limitation relates to the use of PCA for dimensionality reduction. The observed results suggest that, rather than improving the quality of rankings, PCA frequently introduced distortions that weakened the correlation between the derived rankings and HDI. This outcome is likely attributable to the inherent incompatibility between data-driven dimensionality reduction and the theoretically grounded importance of specific socioeconomic indicators. Many critical variables, such as health outcomes, educational attainment, or employment patterns, may not exhibit the highest statistical variance, yet they remain crucial to understanding developmental disparities. Future research could explore an alternative approach where indicators are grouped into predefined thematic categories, such as health, education, and economic indicators, with clustering and ranking processes applied

through each thematic group. Additionally, another unresolved issue pertains to the impact of the number of clusters (k) used in the k-means clustering step. The analysis revealed ambiguous results regarding the optimal number of clusters, and no clear theoretical or empirical guidance emerging from the current configurations. We suggest further research to systematically investigate what is affect the optimal cluster count for socioeconomic rankings.

The overall reliability and generalizability of the rankings are also dependent on the type and quality of the underlying dataset. In particular, if socioeconomic datasets differ significantly in terms of variable coverage, data collection practices, or sampling frames across regions or time periods, the resulting rankings may exhibit non-uniformity. This underscores the need for comprehensive standardization of socioeconomic indicators, ensuring comparability across districts and over time. However, this requirement for standardization immediately encounters the foundational challenge of variable selection itself: what indicators to include, and how to balance theoretically important measures with those that exhibit the strongest empirical variation. Furthermore, a conceptual limitation arises from the reliance on explicit weight assignments at multiple stages of the ranking process, including both the clustering-based ranking and the final alpha-beta weighted combination of cluster and individual distances. This introduces a potential contradiction, as the framework was initially intended to reduce arbitrary weighting decisions by shifting emphasis toward data-driven clustering and ideal point distances. However, the practical results suggest that certain ranking configurations, particularly those without heavy PCA transformation and those using complete datasets without train-test splits, consistently perform well across a broad spectrum of alpha-beta combinations. This indicates that, while the influence of arbitrary weight selection persists, its practical impact may be attenuated when robust clustering and distance calculations are used. Nonetheless, future work could systematically explore how variations in both the clustering weighting and final alpha-beta weighting schemes affect ranking stability and external validity.

6. Appendix

6.1. Clustering Choice

The choice of clustering method was driven by both the nature of the dataset and the specific objectives of this study. DBSCAN was excluded as a viable option due to its inherent property of assigning certain data points as outliers, which contradicts the goals of this analysis.

In the context of constructing socioeconomic rankings, every district needs to be assigned a definitive score, with no district excluded from the final output. DBSCAN performs poorly when clusters exhibit varying densities, a characteristic present in this dataset, as districts differ significantly in population size, economic activity, and other structural features. This sensitivity to density heterogeneity makes DBSCAN unsuitable for ensuring consistent cluster assignments across all districts.

Hierarchical clustering was also excluded due to its computational inefficiency when applied to datasets with high dimensionality. Hierarchical methods become increasingly difficult to manage at larger scales, both in terms of memory requirements and computational time. More importantly, hierarchical clustering is typically useful when the analysis requires understanding nested or hierarchical relationships between clusters, which is unnecessary in this context. Since the objective is to generate distinct, non-overlapping clusters representing districts with broadly similar socioeconomic profiles, a simpler and more scalable method, such as k-means clustering, was deemed more appropriate.

6.2. Handling Missing Values

The treatment of missing data is particularly important in socioeconomic datasets, where missingness is rarely random and often systematically linked to district characteristics. In this dataset, remote or underdeveloped districts tended to have higher rates of missing data, suggesting that the missingness mechanism is "Missing Not At Random" (MNAR). This poses a challenge for traditional complete-case analysis approaches, which would involve excluding any district with incomplete data. Such an approach would not only drastically reduce the sample size but also introduce significant bias, as the excluded districts would disproportionately represent specific socioeconomic and geographic profiles.

To preserve the full sample while maintaining the integrity of inter-district variability, careful consideration was given to imputation strategies. Local spatial or socioeconomic relationships are assumed to be predictive of missing values, k-Nearest Neighbors (kNN) imputation becomes a viable approach. By leveraging the values of the nearest districts, kNN can preserve local patterns of similarity. However, the effectiveness of kNN depends critically on selecting an appropriate number of neighbors (k) and ensuring that highly dissimilar districts do not unduly influence the imputed values.

7. Bibliography

Fadilah, Rizky. "Unsupervised Clustering - Countries Socio-economic Category." RPubs. Accessed February 28, 2025.

Tanmay111999. "Clustering: PCA, K-Means, DBSCAN, Hierarchical." Kaggle. Accessed February 28, 2025.

Han, Sungwon, Donghyun Ahn, Seungeon Lee, Minhyuk Song, Sungwon Park, Sangyoon Park, Jihee Kim, and Meeyoung Cha. "GeoSEE: Regional Socio-Economic Estimation with a Large Language Model." arXiv preprint arXiv:2406.09799 (2024). Accessed February 28, 2025.

World Bank Group. "National Family Survey 2019-2021." Microdata Library. Accessed February 28, 2025.

Chaurasia, Aalok Ranjan. "Human Development in Districts of India, 2019–2021." *Journal of Human Development and Capabilities* (2023). Accessed February 28, 2025.

Ministry of Road Transport and Highways (MORTH), Government of India. *Basic Road Statistics in India: 2018-19.* Accessed February 28, 2025.

International Institute for Population Sciences (IIPS). *National Family Health Survey (NFHS-5), India: Full Report with Final Tables.* Mumbai: IIPS, 2021. Accessed February 28, 2025.

Ministry of Statistics and Programme Implementation (MoSPI), Government of India. *Periodic Labour Force Survey (PLFS), July 2022–June 2023: Household Data (hhv1).* New Delhi: MoSPI, 2024. Accessed February 28, 2025. https://microdata.gov.in/nada43/index.php/catalog/210/datafile/F1.

Ministry of Statistics and Programme Implementation (MoSPI), Government of India. *Periodic Labour Force Survey (PLFS), July 2023–June 2024: Person Data (perv1).* New Delhi: MoSPI, 2024. Accessed February 28, 2025. https://microdata.gov.in/nada43/index.php/catalog/213/datafile/F7.

International Institute for Population Sciences (IIPS) and ICF. *National Family Health Survey (NFHS-5), India, 2019–21.* Mumbai: IIPS, 2021. Accessed February 28, 2025. https://mohfw.gov.in/sites/default/files/NFHS-5_Phase-II_0.pdf.