ORIGINAL RESEARCH ARTICLE

Understanding Fertility Trends of Assam: A District-Level Spatial Analysis Through K-Means Clustering

Chayanika Baruah[1], Ruma Talukdar[2], Saurav Sarma[3]*

[1,2,3]Department of Statistics, Cotton University, Guwahati, India

DOI: 10.55489/njcm.160920255794

ABSTRACT

Background: Understanding regional fertility patterns is crucial for effective demographic planning and policy formulation. This study estimates the district-level Age-Specific Fertility Rate (ASFR) for Assam and employs cluster analysis to classify the districts of Assam based on their current fertility patterns. The cluster analysis identifies distinct groups with similar fertility characteristics, providing insights into regional variations and demographic transitions across the region. The findings highlight significant heterogeneity in fertility levels within Assam, reflecting diverse socio-economic and cultural influences.

Methods: The K-means clustering technique has been used to group the districts of Assam into distinct clusters.

Results: It was found that based on the calculated single-year age-specific fertility rates, the districts of Assam can be divided into two distinct and non-overlapping clusters. A further comparison of some socio-demographic factors between the two clusters revealed that Cluster 1 (Low Fertility Zone) had higher education levels and a greater proportion of Assamese-speaking women compared to Cluster 2 (High Fertility Zone).

Conclusion: The districts of Assam, India, can be divided into two distinct groups with significantly different fertility patterns and demographics of the mothers. These findings can guide targeted reproductive health interventions in high-fertility districts.

Keywords: Age-Specific Fertility Rate, Multivariate Analysis, Cluster Analysis, Demography

ARTICLE INFO

Financial Support: None declared

Conflict of Interest: The authors have declared that no conflict of interests exists.

Received: 30-06-2025, **Accepted**: 04-08-2025, **Published**: 01-09-2025

***Correspondence:** Dr. Saurav Sarma (Email: saurav.sarma@cottonuniversity.ac.in)

How to cite this article: Baruah C, Talukdar R, Sarma S. Understanding Fertility Trends of Assam: A District-Level Spatial Analysis Through K-Means Clustering. Natl J Community Med 2025;16(9):895-906. DOI: 10.55489/njcm.160920255794

www.njcmindia.com | pISSN: 0976-3325 | eISSN: 2229-6816 | Published by Medsci Publications

Introduction

Understanding regional fertility patterns is crucial for effective demographic planning and policy formulation because fertility rates significantly influence household structures, economic planning, resource allocation across families, and the stability of population dynamics over time. India, home to approximately one-fifth of the world's population, stands at the epicentre of a profound demographic transformation characterised by declining fertility rates. The country's total fertility rate (TFR) has halved since 1980 and is now approaching replacement level, marking a significant shift in its population dynamics. However, this national trend masks considerable regional variations, making India one of the most demographically diverse countries in the world. The pace of fertility decline has varied significantly across India's regions, creating a complex demographic landscape.[1] This variation is partly attributed to differences in health systems and adopted policies across Indian states.[2] The interplay between culture, fertility, and policy has shaped this demographic transition, with global trends influencing but not determining the Indian context.[3]

Understanding India's fertility transition requires examining not just the national averages but also the diverse patterns across states and districts, which reflect varying socioeconomic conditions, cultural factors, and policy implementations. These regional differences create a mosaic of demographic profiles within the country, with some states achieving replacement-level fertility while others continue to experience higher rates.[1]

The pace of decline continued through subsequent decades, with the TFR dropping from 5.2 children per woman in 1971-72 to 2.2 by 2015-16.[4] According to the most recent National Family Health Survey (NFHS-5) conducted in 2019-21, India's TFR has further declined to 2.0 children per woman, reaching below the replacement level of 2.1 (Fig 1).[5,6]

While high TFR rates may strain resources and infrastructure, leading to challenges in providing essential services such as healthcare, education, and housing, conversely, declining TFR rates below the replacement level may pose concerns related to ageing populations, workforce shortages, and economic stagnation in the long term.[7]

The TFR of Assam (a northeastern state of India) has also declined rapidly from 3.5 in 1992-93 to 1.87 in 2019-21, which is well below the replacement level of 2.1 per woman and the national TFR rate as well (Fig 1).

Assam, like the other northeastern states of India, is known for its significantly diverse yet distinct geographical, demographic, and socio-economic characteristics that make it different from the rest of India. Hence, a district-level understanding of the current fertility scenario might help us better identify the factors associated.

As per the 2011 Census of India, Assam was administratively divided into 27 districts. However, following the 2011 Census, the Government of Assam initiated further administrative restructuring, resulting in the creation of several new districts. In 2015, five new districts Biswanath (carved out of Sonitpur), Charaideo (from Sivasagar), Hojai (from Nagaon), South Salmara–Mankachar (from Dhubri), and West Karbi Anglong (from Karbi Anglong) were established. As of 2024, Assam comprises 35 districts, reflecting an ongoing process of administrative decentralisation aligned with demographic, geographic, and political considerations. However, the current study is based on 33 districts, as the period of the data used is between 2019 and 2021 (Fig 2). District-level analysis is essential due to Assam's recent administrative decentralization and diverse sociocultural landscape, which likely influence localized fertility patterns.

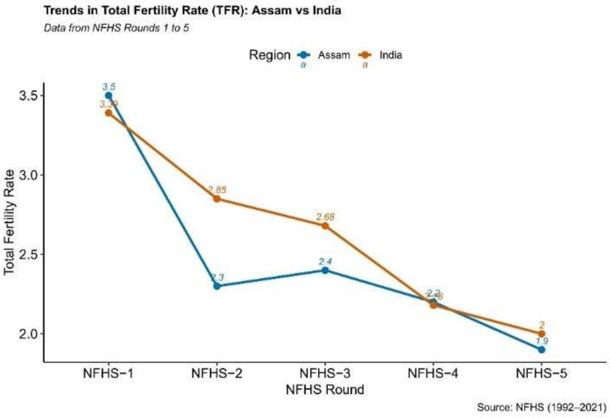


Figure 1: Trend in Fertility Decline of India as well as Assam. From 1992-93 (Round 1) to 2019-21 (Round 5)

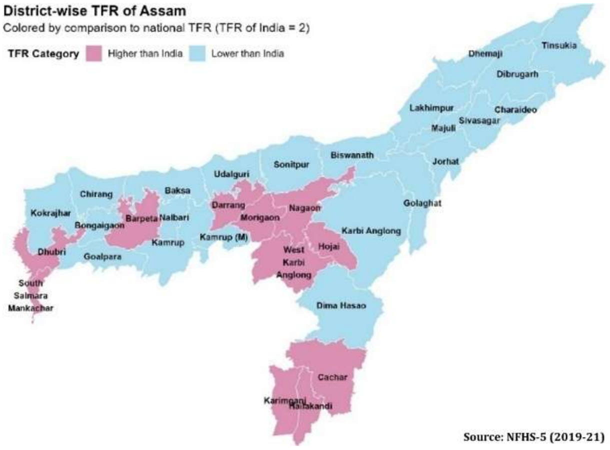


Figure 2: A heatmap of the TFR of all districts of Assam compared to national TFR (2.00) for 2019-21 (Source – NFHS-5)

The socio-economic contrasts among Assam's districts are shaped by geographical, demographic, and developmental factors, with urban centres like Kamrup Metropolitan, Nagaon, and Dubragarh showcasing better infrastructure, education, healthcare, and economic diversification compared to rural or hilly regions such as Karbi Anglong or Dima Hasao. Districts like Goalpara, Morigaon, and Lakhimpur rely heavily on agriculture due to fertile Brahmaputra River lands but face challenges like flood management, while areas like Sonitpur benefit from tourism (e.g., Kaziranga National Park) for additional revenue streams. Literacy rates vary significantly, with districts like Kamrup Metropolitan, Jorhat, and Sivsagar having higher levels than Dhubri, Darrang, and Chirang, which remain below national and state averages. These disparities influence fertility trends: urban areas with higher literacy, economic opportunities, and healthcare access tend to have lower fertility rates due to education-driven family planning and reduced infant mortality, whereas rural and tribal regions often maintain traditional norms favouring larger families due to reliance on agriculture or limited access to services. Overall, socio-economic conditions shaped by education, infrastructure, cultural norms, and policy implementation create a complex landscape that drives varying fertility trends across Assam's districts.[8–11]

This study is an attempt to study the disparity in the fertility curve of all 33 districts of Assam and the socio-demographic contrast and whether they differ significantly among them. However, managing and comparing 33 districts may be cumbersome and unfruitful, as we may fail to discover any hidden pattern among the districts. Therefore, we tried to group them based on similarity of fertility trend for a convenient comparison and interpretability.

Researchers have employed a diverse range of clustering techniques to group Indian states and districts based on various characteristics. Hierarchical clustering, particularly agglomerative hierarchical clustering, represents one of the most straightforward and widely used approaches. This method starts with individual data points as separate clusters and progressively merges with similar clusters until a single large cluster is formed. When applied to Indian contexts, hierarchical clustering has proven effective for analysing COVID-19 patterns[12], nutritional status at district levels[13], and patterns of missing children's cases[14].

K-means clustering, a non-hierarchical partitioning method, is another frequently applied technique for analysing Indian states. This approach has been used to classify districts based on infrastructure development levels[15], analyse COVID-19 data across states and union territories[16], and identify flood risk zones[17]. The K-means method is particularly valued for its ability to produce clear, distinct groupings that can inform policy decisions.

A detailed analysis of age-specific fertility rates (ASFRs) provides a nuanced understanding of how many children people have at different ages within specific regions. This study specifically examines district-level ASFRs in Assam using cluster analysis to identify distinct groups with similar fertility characteristics. We choose K-means clustering for grouping districts of Assam into similar groups because of the nature of the data. The K-means clustering algorithm performs better for numeric data[18] and as we are trying to find similarity of districts based on the single-year ASFR, the K-means is best suited for this study.

Methodology

Source of Data: The National Family Health Survey (Round-5, 2019-2021) dataset for Assam contains information on 34,979 women in the age groups 15-49, out of which 27,215 are ever-married. The sample size of 34,979 women was deemed sufficient for district-level ASFR estimation based on NFHS-5's stratified sampling design, though small sample sizes in some districts may limit precision. For the calculation of Single Year Age-Specific Fertility Rates (ASFRs), the study used de-identified NFHS-5 data, accessed through the DHS Program with ethical approval from the International Institute for Population Sciences (IIPS).[19]

Fertility Rates: The age-specific fertility rate considers the number of children that are being born to k women (typically 1000) in that region during the study period. The sum of each of these ASFRs gives the Total Fertility Rate (TFR).

$$ASFR_x = \frac{Total\ No\ of\ Children\ born\ to\ mother\ of\ age\ x}{Total\ Population\ of\ Women\ age\ x} \times k$$

$$TFR = \sum_{x=15}^{49} ASFR_x$$

The single-year ASFRs were computed for ages 15 to 49 using the ***calc_asfr()*** function from the ***demogsurv*** package within R-Programming.[20] TFR was calculated as the sum of single-year ASFRs for ages 15-49, expressed per woman, following standard demographic practice.[21,22]

Clustering: To identify underlying subgroups within the districts of Assam based on single-year ASFRs, we employed k-means clustering, a widely used unsupervised learning algorithm that partitions data into k mutually exclusive clusters.[23,24] K-means is particularly suited for large datasets with continuous variables and aims to minimise the within-cluster sum of squares (WCSS), thus ensuring that individuals within a cluster exhibit maximal similarity while maximising separation across clusters. Prior to clustering, all input variables were standardised to ensure comparability of scales. To ensure a robust analysis, we addressed issues with incomplete data by implementing multiple imputation for any missing or zero Age-Specific Fertility Rate (ASFR) values, and outliers were assessed using z-scores to ensure robust clustering.

Before applying the k-means clustering algorithm, an essential preliminary step is to determine the appropriate number of clusters to be used. The literature offers a variety of methods for estimating this number, each grounded in different theoretical considerations. One widely used technique is the *Elbow Method*, which evaluates the total intra-cluster variation typically measured as the within-cluster sum of squares (WSS) as a function of the number of clusters. The optimal number of clusters is identified at the "elbow" point, where the rate of decrease in WSS begins to level off.[25] Another approach is the *Average Silhouette Method*[26], which calculates the average silhouette width for different values of k. The silhouette value measures how similar an object is to its own cluster compared to other clusters, and the optimal number of clusters is the one that maximises this average. A third method, the *Gap Statistic*, compares the observed WSS with that expected under a reference null distribution. The optimal number of clusters is the value of k that yields the largest gap between the observed and expected values, indicating a significant clustering structure.[27]

The optimal number of clusters was determined using a combination of the elbow method, silhouette analysis, and the Krzanowski and Lai (KL) Index[28], reflecting a balance between cluster homogeneity and dispersion. This approach enables a data-driven identification of homogeneous subgroups, facilitating further comparative analysis across clusters.

The KL index proposed by Krzanowski and Lai is defined as

$$KL_{(q)} = \left| \frac{DIFF_q}{DIFF_{q+1}} \right|$$

Where, $DIFF_q = (q-1)^{2/p}\ \mathrm{trace}(W_{q-1}) - q^{2/q}\ \mathrm{trace}(W_q)$

The value of q, maximizing $KL_{(q)}$ is regarded as specifying the optimal number of clusters[29].

The optimal number of clusters was selected by triangulating results from the Elbow Method, Silhouette Analysis, and KL Index, prioritizing the point of convergence across these metrics. The NbClust package (version 3.0.1) was used with default settings, including distance metric = Euclidean, minimum clusters (min.nc) =2, and method = "kmeans", to find the optimum number of clusters for our data. The package provides nearly 30 different indices to find the optimum number of clusters including the KL index, Silhouette, Gap etc.[30]

Results

Fertility Rates: We calculated the single-year ASFRs for India, all northeastern states of India, which include Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, and Tripura, and for all 33 districts of Assam from the individual record data of the fifth round of NFHS [see Table 1]. While Manipur and Meghalaya have higher TFRs than the national TFR (TFR of India = 2.0), the remaining northeastern states showed a lower TFR level, with Sikkim being the lowest. In terms of district-level fertility, Barpeta, Cachar, Chirang, Darrang, Dhubri, Hailakandi, Hojai, Karbi Anglong, Karimganj, Kokrajhar, Morigaon, Nagaon, South Salmara Mancarhar, and West Karbi Anglong showed higher TFR compared to the state TFR (TFR of Assam = 1.87), and the remaining were less than the state TFR (Fig 3).

A comparison of the fertility pattern of Assam with that of India reveals that Assam has a slightly higher fertility in the 15-19 and 30-39 age ranges compared to the national level. However, fertility is substantially lower in the 20-30 age range compared to the national ASFR (Fig 4) [Table 1].

Figure 5 depicts the age-specific fertility rate (ASFR) per 1000 women across all districts of Assam, based on calculated single-year ASFR. The vertical axis represents the age of women (15–49 years), while the horizontal axis denotes the individual districts. The intensity of the colour gradient corresponds to the magnitude of the ASFR, with deeper shades of pink indicating higher fertility rates. Additionally, asterisks (*) mark the modal age (i.e., the age group with the highest ASFR) for each district. The heatmap visualization reveals several notable patterns. First, fertility rates tend to peak within the age range of 20 to 25 years across most districts, consistent with demographic patterns typical of early childbearing populations. The mode (maximum ASFR) is predominantly observed at ages 21, 22, or 23 in most districts, signifying the concentration of childbearing during early adulthood.

Districts such as Dhubri, Barpeta, and Goalpara exhibit comparatively higher ASFR values, as evidenced by more intense colour shades, particularly in the early twenties. In contrast, districts like Kamrup Metropolitan and Jorhat demonstrate lighter shades throughout, indicating relatively lower fertility levels. The presence of the mode at younger ages (e.g., 19–20 years) in districts like Dhubri and South Salmara Mankachar may reflect earlier initiation of childbearing in these regions, potentially influenced by sociocultural factors.

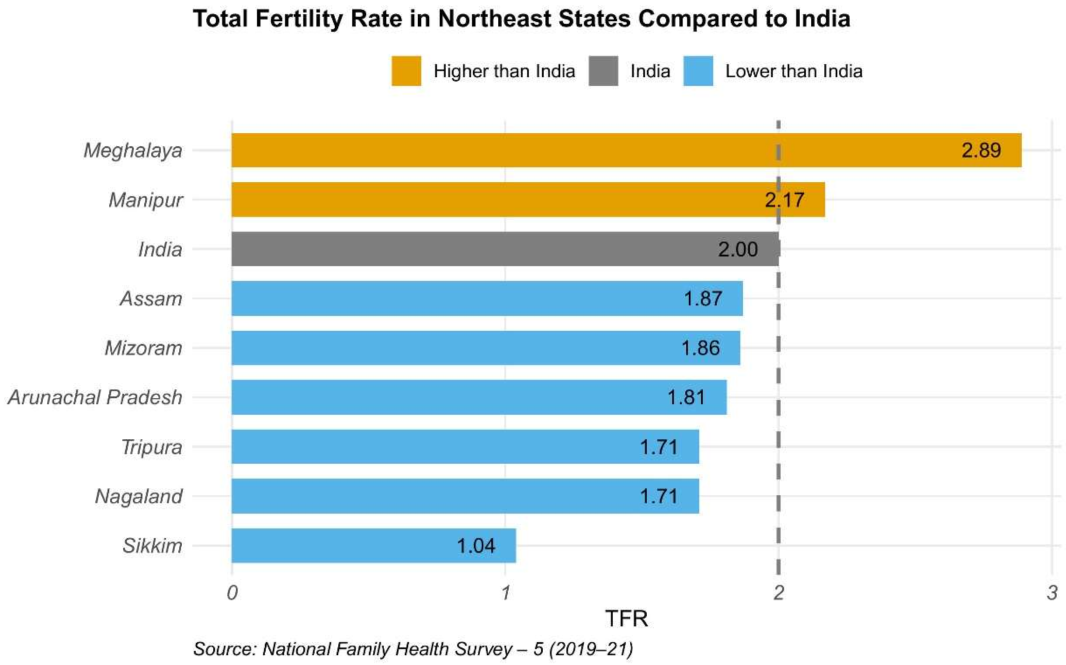
Furthermore, ASFR declines substantially after the mid-twenties across all districts, with minimal fertility contribution observed after 35 years of age. Zero ASFR values (represented by white shades in Fig 5) in older age groups likely reflect small sample sizes within certain districts, potentially affecting Total Fertility Rate (TFR) estimates.

District Level Clustering: To determine whether the data had a meaningful cluster structure (i.e., non-randomness), the Hopkins statistic was calculated. The value of the statistic was found as, H = 0.81, which is significantly greater than the 0.5 threshold, suggesting a high likelihood of grouping and hence supporting the use of a clustering algorithm.[31,32]

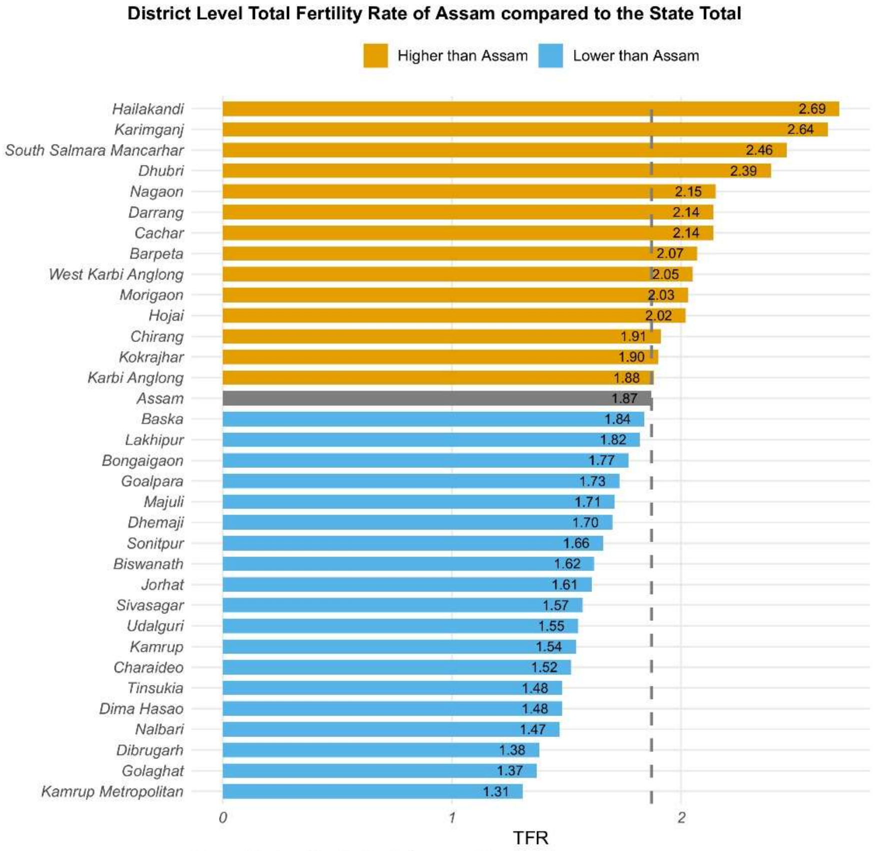
Table 1: Single Year ASFR (per 1000 women) of India and the North-Eastern States

Age (in years)	India	Assam	Arunachal Pradesh	Manipur	Meghalaya	Mizoram	Nagaland	Sikkim	Tripura
15	2.84	9.85	4.33	5.4	4.52	0	0	0	28.51
16	10.38	20.14	8.96	24.69	14.15	5.78	2.84	7.33	52.23
17	29.13	64.94	32.07	43.94	39.05	26.48	10.82	3.92	104.52
18	64.86	89.73	61.51	62.86	78.32	41.24	34.86	54.33	133.07
19	109.17	117.31	83.21	76.73	113.56	43.75	41.42	36.94	135.55
20	142.93	133.9	92.18	97.32	149.94	83.01	69.79	65.39	152.77
21	163.11	135.41	96.67	119.31	140.2	80.25	80.07	45.33	134.79
22	175.69	142.37	116.61	92.84	155.14	71.07	96.55	81.06	111.52
23	174.43	134.58	111.79	106.81	128.54	120.92	95.09	53.14	77.14
24	168.97	128.19	112.47	131.44	144.63	112.11	111.81	41.19	118.07
25	155.67	122.3	109.58	126.79	146.33	115.84	83.86	54.77	112.06
26	141.12	102.95	108.52	93.32	142.29	119.07	113.02	71.79	81.7
27	118.81	100.1	101.86	133.46	154.79	97.12	118.7	93.46	78.69
28	103.75	94.84	87.25	129.24	135.32	85.69	92.45	74.02	64.08
29	84.59	75.82	100.74	124.03	128.55	97.39	116.17	52.71	52.93
30	74.17	74.52	82.68	126.78	155.86	100.14	82.27	44.16	65.29
31	60.11	62.52	70.88	107.18	136.42	101.69	89.79	23.76	50.31
32	47.29	52.95	74.74	103.27	99.42	75.92	76.74	73.61	40.68
33	38.26	43.09	61.26	73.3	136.71	72.49	89.43	36.5	29.13
34	30.67	36.09	63.42	87.52	107.57	95.98	68.3	42.07	21.84
35	22.8	37.55	42.41	93.11	101.4	78.32	52.92	20.78	9.15
36	17.49	23.77	44.21	63.99	78.82	60.94	65.58	18.2	20.6
37	12.99	15.4	32.2	50.19	80.92	42.23	17.05	26.62	2.15
38	10.37	19.72	13.48	35.34	69.69	43.31	27.03	3.81	8.25
39	6.76	7.1	16.92	26.97	48.95	33.13	16.91	4.19	10.53
40	5.57	4.7	16.84	13.47	46.7	16.35	19.44	6.57	2.87
41	3.65	6.39	5.42	3.52	47.84	8.82	19.19	3.48	2.03
42	2.55	2.87	7.61	8.49	31.75	20.83	5.47	0	2.21
43	1.68	5.56	6.92	3.78	21.7	4.87	6.44	0	2.1
44	1.79	1.48	10.44	2.09	14.42	3.32	2.33	5.52	0
45	1.17	1.78	4.07	1.91	20.78	1.64	0	0	0
46	0.33	0	4.78	1.53	2.27	0	3.08	0	0
47	0.55	0	5.48	0.67	10.78	0	0	0	1
48	0.3	0	1.25	0	3.21	0	0	0	0
49	0.46	0	17.67	0	0	0	0	0	0
TFR	**2.00**	**1.87**	**1.81**	**2.17**	**2.89**	**1.86**	**1.71**	**1.04**	**1.71**

Source: National Family Health Survey 5 (2019-2021)



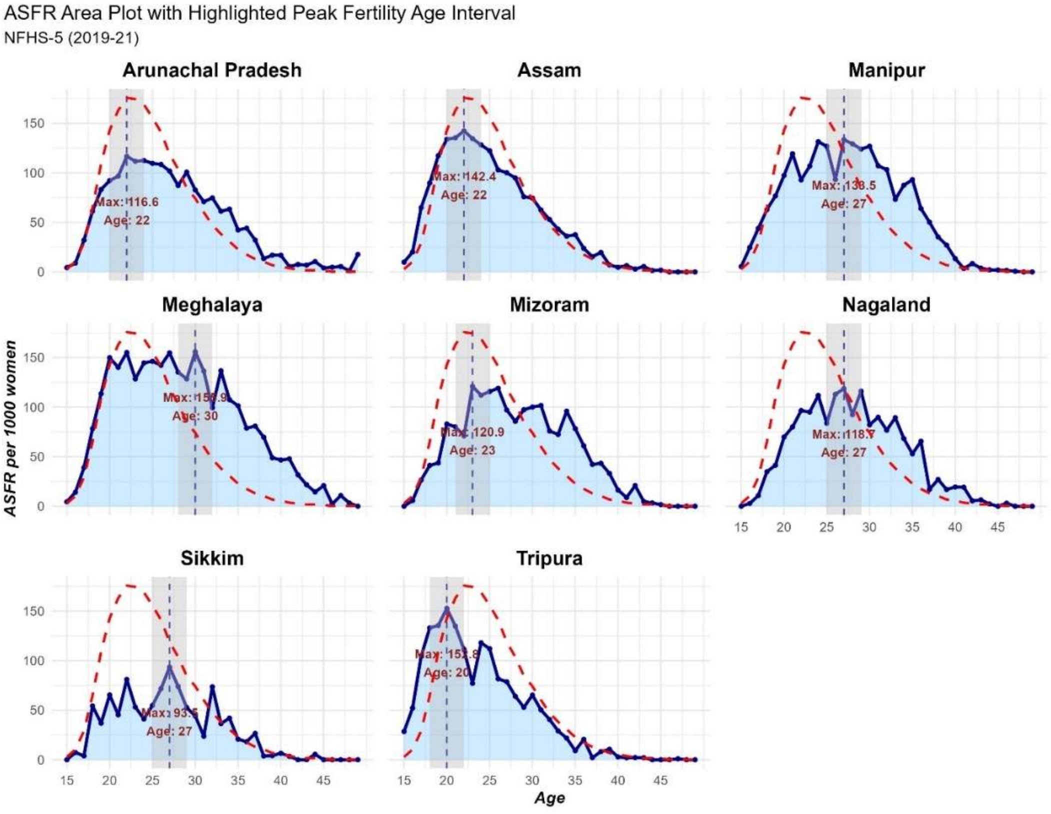
3(a) Comparison of northeastern states with national TFR



Source: National Family Health Survey – 5 (2019–21)

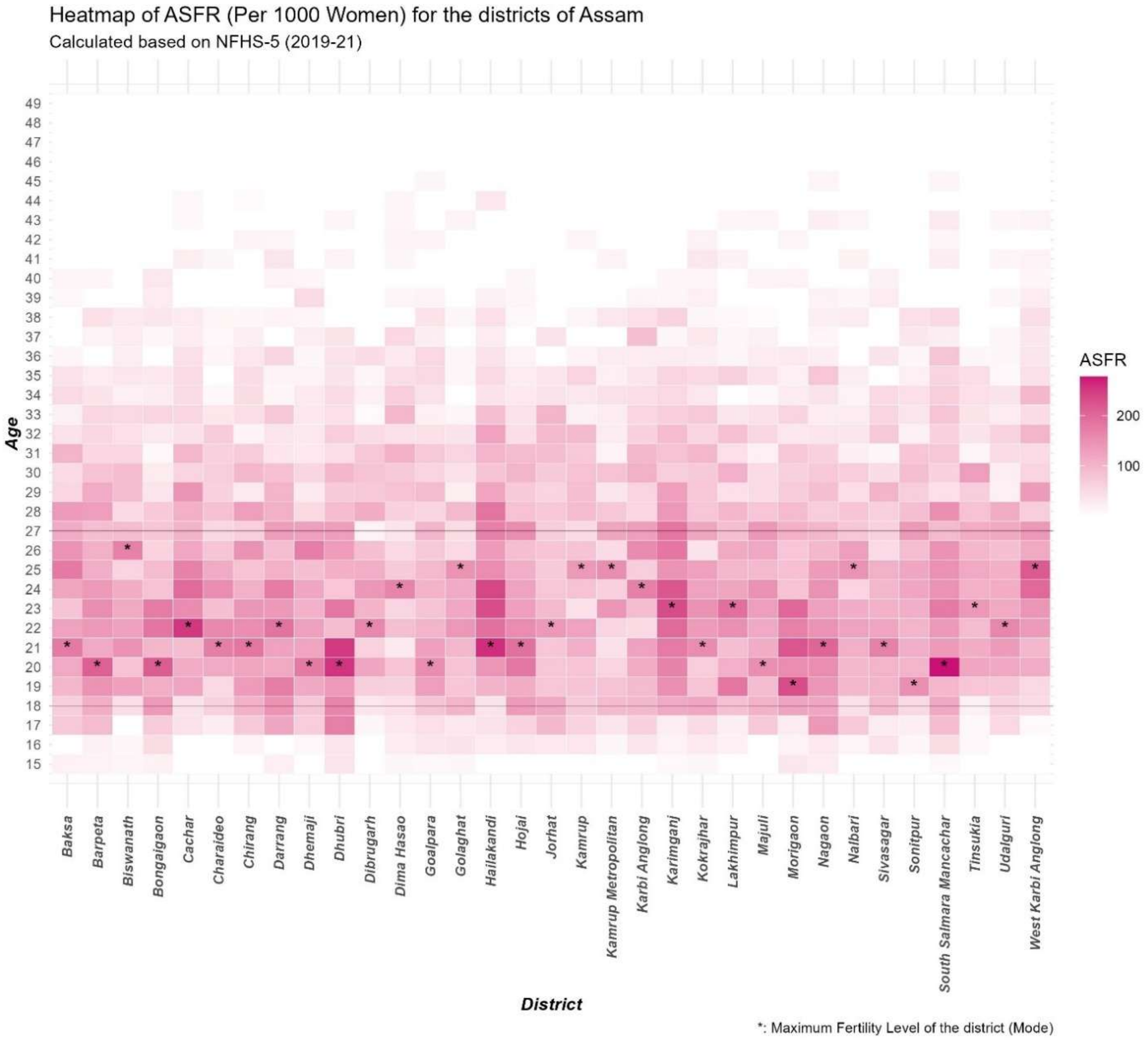
3(b) Comparison of districts with State TFR

Figure 3: Total Fertility Rate of Northeastern States and Districts of Assam



Red dashed line = India reference,
Vertical Dashed line = modal age, grey band = modal window

Fig 4: An area plot of the single year ASFR of Northeastern states in comparison to India, calculated from NFHS-5 (2019-21). The red dashed line presents the fertility curve of India. The vertical black dashed lines represent the modal age of fertility with text indicating the modal age and highest fertility per 1000 women for the state.



*Indicates the highest fertility rate (modal age of fertility curve).

Figure 5: A heatmap for the ASFR's (per 1000 women) of all the districts of Assam calculated from NFHS-5 data

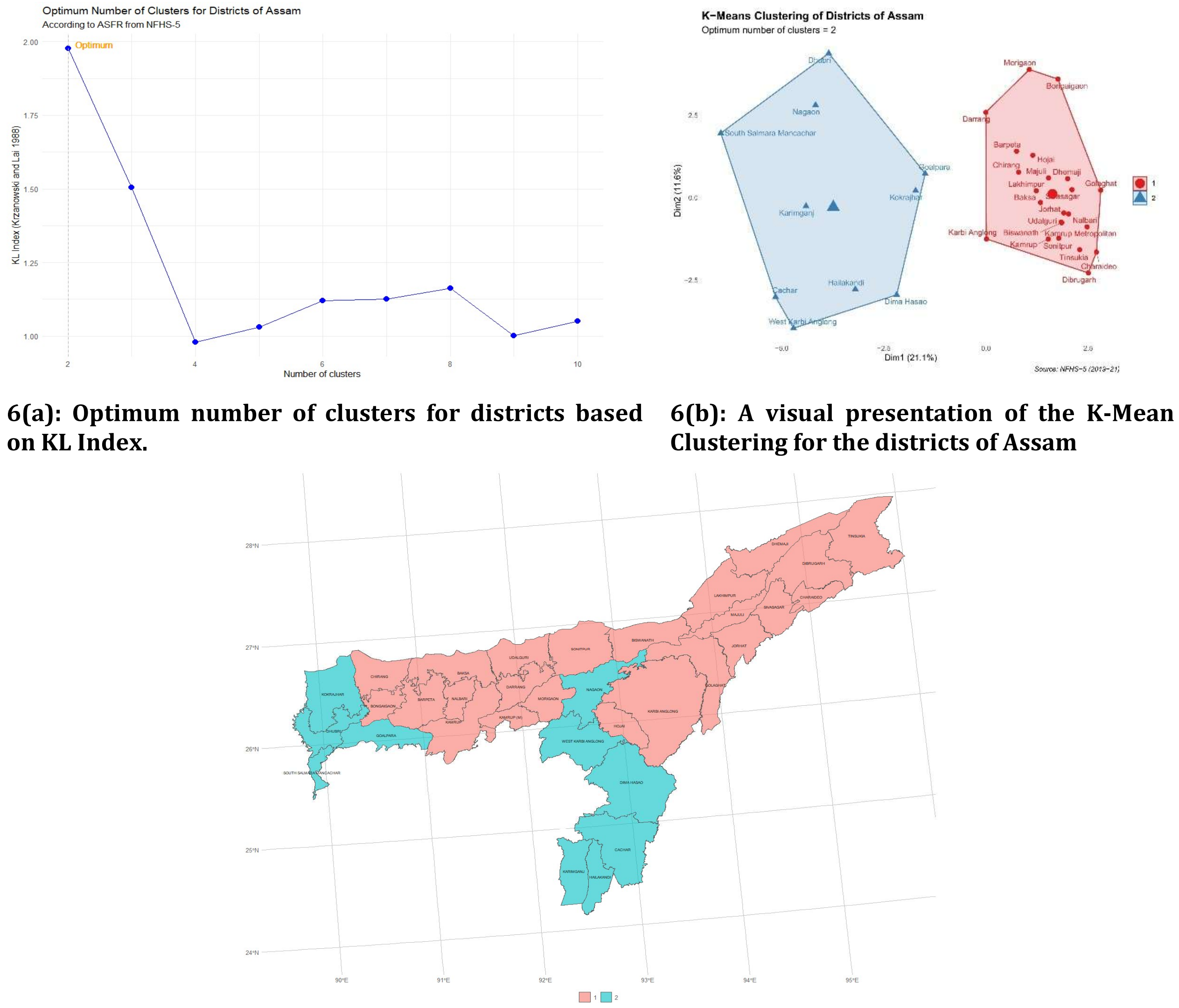
A visual inspection using **"silhouette"** (for average silhouette width), **"wss"** (for total within sum of square), and **"gap_stat"** (for gap statistics) along with the KL Index (Fig 6a) suggests that two clusters are optimum to capture the distinct fertility pattern of the districts of Assam. Accordingly, K-mean clustering is performed in R with two centres. The spherical clusters are found to be non-overlapping, indicating maximum between cluster distance and minimal within cluster distance (Fig 6b). Sensitivity analyses using hierarchical clustering confirmed the robustness of the two-cluster solution, with similar district groupings (results available upon request).

The 33 districts of Assam are divided into two groups with Cluster 1 consisting of 24,142 members and Cluster 2 containing 10,837 women [Table 2] (Fig 6b). Cluster 1 has a TFR of 1.67 and Cluster 2 has a TFR of 2.23. Also, Cluster 2 has consistently higher fertility trend than Cluster 1 and the state fertility rates (Fig: 7). Therefore, it is fair to call Cluster 1 as ***"Low Fertility Zone"*** and Cluster 2 as ***"High Fertility Zone"***.

Cluster 1 exhibits lower ASFRs across all age groups, with a notable peak at ages 20-24 (120.4 per 1000 women) with modal age of fertility at 22 years, while Cluster 2 shows higher ASFRs, peaking at ages 20-24 (162.4 per 1000 women) with a modal age of fertility at 21 years, indicating delayed fertility in Cluster 1.

Table 2: Member districts in the two clusters

Clusters	Districts
Cluster 1	Baksa, Barpeta, Biswanath, Bongaigaon, Charaideo, Chirang, Darrang, Dhemaji, Dibrugarh, Golaghat, Hojai, Jorhat, Kamrup, Kamrup Metropolitan, Karbi Anglong, Lakhimpur, Majuli, Morigaon, Nalbari, Sivasagar, Sonitpur, Tinsukia, Udalguri
Cluster 2	Cachar, Dhubri, Dima Hasao, Goalpara, Hailakandi, Karimganj, Kokrajhar, Nagaon, South Salmara Mancachar, West Karbi Anglong

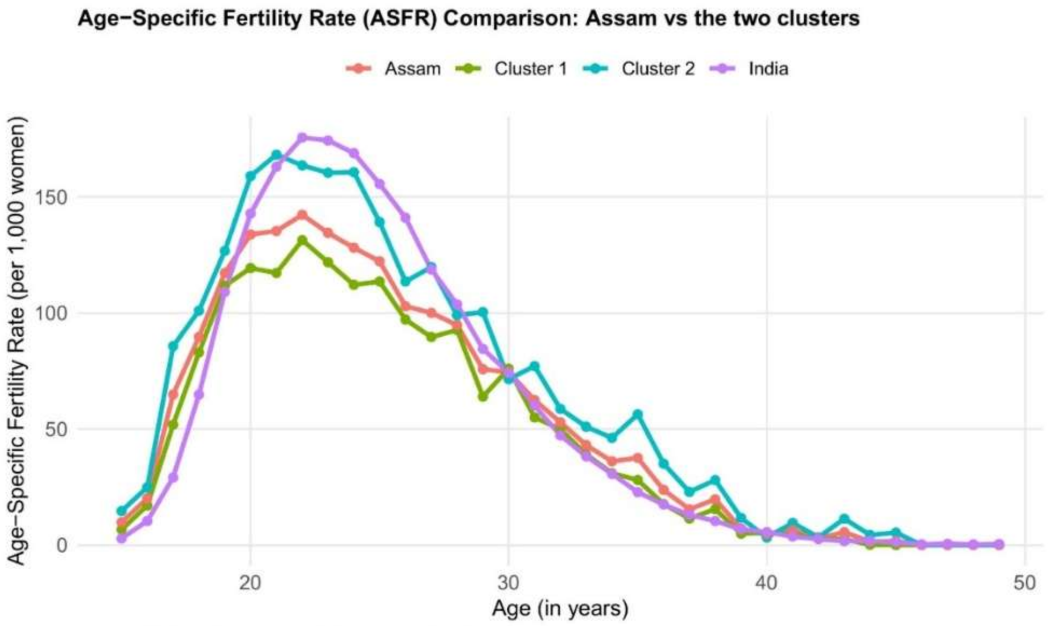


6(a): Optimum number of clusters for districts based on KL Index.

6(b): A visual presentation of the K-Mean Clustering for the districts of Assam

6(c): A Spatial presentation of the two clusters of districts

Figure 6: Cluster Analysis of districts of Assam



Source: National Family Health Survey – 5 (2019–21)

Figure 7: A comparison of the ASFR pattern between Assam and the two clusters of districts with that of India

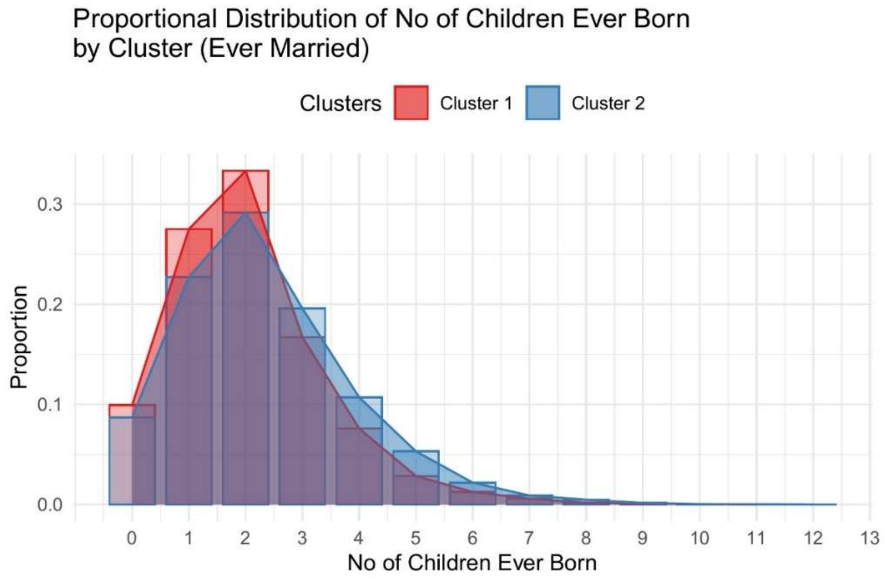


Figure 8: An area plot comparing the distribution of number of children ever born between the two cluster of districts

Between Cluster Comparison: We performed some Pearson's Chi Square test of independence to compare a few socio - demographic characteristics, which may affect the fertility of the women in the two clusters for a better characterization of the two clusters. Chi-square tests assume independence of observations, which may be violated due to geographic clustering. Future studies should validate findings with larger samples. Initially a median test using the Wilcoxon signed rank test was performed to compare the total children ever born and the total number of living children of mothers in the two clusters but since the median and inter-quartile range (IQR) are identical we opted for the Yuen's robust t-test for trimmed mean.

A univariate binary logistic model was run to calculate the odds of being a member of Cluster 1 compared to Cluster 2. An odds ratio greater than 1 indicates that the category has more change of being in Cluster 1 compared to Cluster 2, while a value far below 1 indicates that the category is more prominent in Cluster 2 [Table 3].

Discussion

A Geo-spatial visualization of the districts in the two clusters showed a distinct pattern. Districts in a particular belt or region are found to be members of the same cluster. While cluster 1 of districts are in a continuous range of area, cluster 2 of districts are in two different but neighbouring ranges of area (Fig 6c). The geographic clustering of districts in Cluster 2 (e.g., Cachar, Dhubri) may reflect cultural practices favouring larger families or limited healthcare access, consistent with Deka & Sarma (2022).[10] A comparison of the single year ASFR of the two clusters with that of state fertility trend reveals that the cluster 2 of districts have consistently higher fertility rates throughout the fertility period (Fig: 7) with a peak fertility trend in the 20 - 25 age range.

A further investigation revealed some key difference between the two clusters. A significantly higher proportion of women in cluster 1 (Low Fertility Zone) has attained higher education compared to cluster 2 (High Fertility Zone) (OR: 1.75, 95% CI: 1.58 - 1.95, p<0.001), whereas, the proportion of woman with primary education is higher in cluster 2 (OR: 0.88, 95% CI: 0.81 - 0.95, p<0.001). A significantly higher proportion (77%) of members in Cluster 1 speak the state language (Assamese), while Cluster 2 has equal proportion of Assamese and Bengali speaking members (40.5% and 40.8% respectively). The Cluster 1 has a higher proportion of Hindu population (73.9%) while Cluster 2 has an almost equal proportion of Hindu and Muslim members (45.7% and 47.8% respectively). Cluster 1 has a significant proportion of caste members (62.8%) while Cluster 2 has more members that belong to neither any caste nor tribe (46.9%). While it comes to physical fitness of the women, Cluster 2 has higher proportion of non-anaemic and normal BMI level women (41.34% and 72.34% respectively) compared to Cluster 1 (31.55% and 65.89% respectively).

The median number of children ever born to the mothers in both clusters are 2 with an IQR of (1-3), however, the concentration around the median differs significantly as evident by a median test (Wilcoxon Sign Ranked test: p <0.001). Despite similar medians and IQRs across groups, distributional asymmetries likely contributed to this result. A Yuen's robust t-test comparing 20% trimmed means of total number of children ever born across the two clusters (restricted to ever-married individuals) revealed a statistically significant difference, t (9419.79) = 14.87, p <0.001. The estimated trimmed mean difference was 0.27 (95% CI: 0.31 to 0.23), with an explanatory effect size (ξ) of 0.15, indicating a small but meaningful effect.

Table 3: Comparison of some key socio-demographic characteristic of the women in the two clusters

Demographic Characteristic	N	Assam	Cluster 1 N = 24,142	Cluster 2 N = 10,837	TFR	OR (95% CI)[$] (Cluster 1 vs. Cluster 2)	Between Cluster Comparison
Highest educational level	34,979						
No education		6,515 (18.63%)	4,442 (18.40%)	2,073 (19.13%)	2.27		**161 (<0.001)[1]**
Primary		4,700 (13.44%)	3,067 (12.70%)	1,633 (15.07%)	2.29	0.88 (0.81-0.95)***	
Secondary		21,021 (60.10%)	14,467 (59.92%)	6,554 (60.48%)	1.87	1.03 (0.97-1.09)	
Higher		2,743 (7.84%)	2,166 (8.97%)	577 (5.32%)	1.31	1.75 (1.58-1.95)***	
Type of place of residence	34,979						
Urban		4,291 (12.27%)	3,062 (12.68%)	1,229 (11.34%)	1.5		**12.5 (<0.001)[1]**
Rural		30,688 (87.73%)	21,080 (87.32%)	9,608 (88.66%)	1.93	0.88 (0.82-0.94)***	
Native language of respondent	34,979						
Assamese		22,976 (65.69%)	18,590 (77.00%)	4,386 (40.47%)	1.75		**5793 (<0.001)[1]**
Bengali		6,498 (18.58%)	2,073 (8.59%)	4,425 (40.83%)	2.34	0.11 (0.10-0.12)***	
Other		5,505 (15.74%)	3,479 (14.41%)	2,026 (18.70%)	1.63	0.41 (0.38-0.43)***	
Religion	34,979						
Hindu		22,782 (65.13%)	17,831 (73.86%)	4,951 (45.69%)	1.58		**2623 (<0.001)[1]**
Muslim		10,621 (30.36%)	5,445 (22.55%)	5,176 (47.76%)	2.38	0.29 (0.28-0.31)***	
Others		1,576 (4.51%)	866 (3.59%)	710 (6.55%)	1.47	0.34 (0.31-0.38)***	
Ethnicity	34,979						
Caste		19,311 (55.21%)	15,162 (62.80%)	4,149 (38.29%)	1.65		**1992 (<0.001)[1]**
Tribe		4,589 (13.12%)	2,977 (12.33%)	1,612 (14.87%)	1.56	0.51 (0.47-0.54)***	
Others		11,079 (31.67%)	6,003 (24.87%)	5,076 (46.84%)	2.26	0.32 (0.31-0.34)***	
Wealth index combined	34,979						
Poorest		13,014 (37.21%)	8,381 (34.72%)	4,633 (42.75%)	2.28		**388 (<0.001)[1]**
Poorer		11,692 (33.43%)	8,093 (33.52%)	3,599 (33.21%)	1.81	1.24 (1.18-1.31)***	
Middle		5,971 (17.07%)	4,263 (17.66%)	1,708 (15.76%)	1.49	1.38 (1.29-1.47)***	
Richer		3,160 (9.03%)	2,421 (10.03%)	739 (6.82%)	1.38	1.81 (1.66-1.98)***	
Richest		1,142 (3.26%)	984 (4.08%)	158 (1.46%)	1.18	3.44 (2.91-4.10)***	
Anaemia level	26,464						
Severe		577 (2.18%)	447 (2.45%)	130 (1.58%)	1.46		**283 (<0.001)[1]**
Moderate		9,193 (34.74%)	6,783 (37.22%)	2,410 (29.25%)	2.09	0.79 (0.66-0.94)**	
Mild		7,538 (28.48%)	5,245 (28.78%)	2,293 (27.83%)	1.86	0.67 (0.56-0.80)***	
Not anaemic		9,156 (34.60%)	5,750 (31.55%)	3,406 (41.34%)	1.71	0.49 (0.41-0.59)***	
Current contraceptive method	27215[#]						
Not using		10,969 (40.30%)	7,768 (41.28%)	3,201 (38.11%)	1.54		**24.3 (<0.001)[1]**
Using Modern or Traditional Methods		16,246 (59.70%)	11,048 (58.72%)	5,198 (61.89%)	2.84	0.92 (0.88-0.96)***	
BMI Level	34,311						
Underweight		5,823 (16.97%)	4,105 (17.37%)	1,718 (16.10%)	2.2		**207 (<0.001)[1]**
Normal		23,295 (67.89%)	15,574 (65.89%)	7,721 (72.34%)	1.92	0.84 (0.79-0.90)***	
Overweight		4,369 (12.73%)	3,271 (13.84%)	1,098 (10.29%)	1.35	1.25 (1.14-1.36)***	
Obese		824 (2.40%)	688 (2.91%)	136 (1.27%)	1.17	2.12 (1.76-2.58)***	
Total Children Ever Born, Median (IQR)	27215[#]	2 (1, 3)	2 (1, 3)	2 (1, 3)			**14.87 (<0.001)[2]**
Total No of Living Children, Median (IQR)	27215[#]	2 (1, 3)	2 (1, 3)	2 (1, 3)			**15.39 (<0.001)[2]**

[1]Pearson's Chi-squared test; [2]Yuen's test for trimmed means; [#]Ever married sample; [$]Cluster 2 as reference category

Odds ratios were calculated with Cluster 2 (High Fertility Zone) as the reference category, where an OR > 1 indicates a higher likelihood of belonging to Cluster 1 (Low Fertility Zone).

A closer inspection reveals that the distribution of number of children born is left skewed from the median for cluster 1 and right skewed for cluster 2. While cluster 1 has a higher proportion of 0-2 children, cluster 2 has higher proportion of 3-6 children (Fig: 8). However, the differences in fertility may also be influenced by unmeasured factors such as access to contraception or healthcare services, which warrant further investigation.

Our findings align with Joshi et al. (2021)[1], who noted spatial variations in contraceptive use across Indian districts, suggesting that lower fertility in Cluster 1 may be linked to higher contraceptive prevalence.

The analysis of the two distinct clusters reveals significant socio-demographic differences that have important implications for policy development tailored to each cluster's specific needs. In particular, Cluster 1 (Low Fertility Zone) is characterized by higher educational attainment among women and a predominant state language-speaking population with a majority being Hindu caste members. This suggests policies in these districts might focus on maintaining or enhancing the existing educational infrastructure and supporting cultural preservation initiatives that respect the linguistic and religious demographics.

Cluster 2 (High Fertility Zone) presents unique challenges and opportunities due to its higher fertility rates, lower educational attainment among women, a more diverse language composition with significant Bengali speakers, and balanced Hindu-Muslim populations. For these districts, policy interventions could be specifically designed to address the drivers of high fertility. This includes enhancing female education as it is strongly associated with reduced fertility rates; initiatives could involve scholarships for girls' education or adult literacy programs aimed at women already out of school. Moreover, improving access to family planning services in Cluster 2 can empower women with more control over their reproductive choices, potentially leading to a decrease in the high fertility trends observed.

Additionally, considering the physical health aspects revealed by the higher proportion of non-anaemic and normal BMI level individuals in Cluster 2 compared to Cluster 1, policies could also incorporate nutritional programs targeted at improving overall community health. Such comprehensive strategies would not only address immediate demographic concerns but also contribute to long-term socio-economic development goals for these high fertility areas.

In essence, adopting zone-specific policies that consider the unique characteristics of each cluster will be crucial in effectively managing and supporting their developmental trajectories. These tailored approaches can ensure more equitable resource distribution and maximize the positive impact on community well-being across both clusters.

However, it should be noted that NFHS-5 data may be subject to recall bias in fertility reporting, and incomplete coverage in remote districts could affect ASFR estimates. These limitations should be considered when interpreting results.

Conclusion

This study identified two distinct fertility zones in Assam, with Cluster 1 (Low Fertility Zone) characterized by higher education and Assamese-speaking populations, and Cluster 2 (High Fertility Zone) marked by lower education and a higher proportion of Bengali-speaking and Muslim populations. 11 out of 33 districts were found to have higher TFR than the national TFR (2.0). The data comprised of mostly rural dweller (87.73%), with one third of the sampled population being Hindu and Assamese speaking. More than 70% of the sampled women are categorized as either poorest or poor and a high percentage (65.40%) were found to be anaemic. A cluster analysis suggested a 2-cluster division with 23 districts identified as “Low Fertility Zone” and 10 districts were identified as “High Fertility Zone”. The two zones differ significantly in terms of the education attainment, religion, caste, physical fitness etc. of the mothers. This study is a humble attempt to geospatial grouping based on human fertility; further analysis would help identifying more key factors that are contributing to these differences so that more zone-specific policies can be adopted to restrict the fertility patters of the region in an optimal level. Future research should explore longitudinal trends in ASFRs and incorporate additional variables, such as contraceptive access and maternal healthcare utilization, to elucidate causal pathways.

Acknowledgement: We gratefully acknowledge the Demographic and Health Surveys (DHS) Program and the International Institute for Population Sciences (IIPS), Mumbai, for granting access to the dataset used in this study.

Individual Authors’ Contributions: CB: Analysis and Writing of the article, RT: Manuscript editing and review, SS: Analysis of data in R programming.

Availability of Data: The NFHS data can be downloaded after approval from the DHS Program (https://dhsprogram.com/data/available-datasets.cfm)

No use of generative AI tools This article was prepared without the use of generative AI tools for content creation, analysis, or data generation. All findings and interpretations are based solely on the authors' independent work and expertise.

References

1. Joshi S, Uttamacharya, Borkotoky K, Gautam A, Datta N, Achyut

P, Nanda P, Verma R. Spatial Variation in Contraceptive Practice Across the Districts of India, 1998-2016. Spat Demogr. 2021;9(2):241-269. DOI: https://doi.org/10.1007/s40980-021-00092-9 PMid:34722854 PMCid:PMC8549954

2. Munshi V, Yamey G, Verguet S. Trends in state-level child mortality, maternal mortality, and fertility rates in India. Health Aff (Millwood). 2016;35(10):1759-1763. DOI: https://doi.org/10.1377/hlthaff.2016.0552 PMid:27702946
3. Nayana V, Sandeep J. From Boom to Bust: Unpacking India's Fertility Decline. Indian J Res Anthropol. 2024;10(1):79-94. DOI: https://doi.org/10.21088/ijra.2454.9118.10124.9
4. Halli SS, et al. Fertility and family planning in Uttar Pradesh, India: major progress and persistent gaps. Reprod Health. 2019;16(1):1-10. DOI: https://doi.org/10.1186/s12978-019-0790-x PMid:31443724 PMCid:PMC6706892
5. Visaria L. India's date with second demographic transition. China Popul Dev Stud. 2022;6(3):316-337. DOI: https://doi.org/10.1007/s42379-022-00117-w
6. Tiwari AK, Maurya RK, Singh PK. Proximate Determinant of Fertility in India and Estimation of Total Fertility Rate. Braz J Biom. 2025;43(1):1-15. DOI: https://doi.org/10.28951/bjb.v43i3.771
7. Bhattacharjee NV, et al. Global fertility in 204 countries and territories, 1950-2021, with forecasts to 2100: a comprehensive demographic analysis for the Global Burden of Disease Study 2021. Lancet. 2024;403(10440):2057-2099. DOI: https://doi.org/10.1016/S0140-6736(24)00550-6
8. Deluwar H, Hazarika C. An empirical analysis of income and livelihood pattern in sandbar areas along the river Brahmaputra. Int J Soc Sci Humanit. 2020;10(2):35-41. DOI: https://doi.org/10.18178/ijssh.2020.V10.1010
9. Census Commission of India. District-wise basic data - Assam. New Delhi: Government of India; 2011. Available from: https://assam.census.gov.in/basicdata.php [Accessed on June 04, 2025].
10. Deka M, Sarma S. A quantitative study of the maternal health care of different districts of Assam, India. Int. Res. J. Social Sci. 2022;11(1):13-21.
11. Baruah S, Borah MC. Inter-district disparities in industrial growth of Assam. Int J Approx Reason. 2017;5(4):1027-1034. DOI: https://doi.org/10.21474/IJAR01/3601
12. Chakraborty S. Monitoring COVID-19 cases and vaccination in Indian states and union territories using unsupervised machine learning algorithm. Ann Data Sci. 2022;10(5):967-989. DOI: https://doi.org/10.1007/s40745-022-00404-w
13. Striessnig E, Bora JK. Under-five child growth and nutrition status: spatial clustering of Indian districts. Spat Demogr. 2020;8(1):63-84. DOI: https://doi.org/10.1007/s40980-020-00058-3
14. J T, Kumar S, Panda PS, et al. Geospatial hotspot analysis and endemicity trends of missing and unrecovered children in India. Cureus. 2023;15(6):e39955. DOI: https://doi.org/10.7759/cureus.39955 PMID: 37416019; PMCID: PMC10319941.
15. Dube M, Yadav SK, Singh V. Uncovering regional disparities in infrastructural development of Uttar Pradesh: an exploratory factor analysis. J Reliab Stat Stud. 2022;15(1):21-36. DOI: https://doi.org/10.13052/jrss0974-8024.1512
16. Kumar S. Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. Ann Data Sci. 2020;7(3):417-425. DOI: https://doi.org/10.1007/s40745-020-00289-7
17. Matheswaran K, Alahacoon N, Pandey RK, Amarnath G. Flood risk assessment in South Asia to prioritize flood index insurance applications in Bihar, India. Geomat Nat Hazards Risk. 2018;10(1):26-48. DOI: https://doi.org/10.1080/19 475705.2018.1500495
18. Adolfsson A, Ackerman M, Brownstein NC. To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recognit. 2019;88(1):13-26. DOI: https://doi.org/10.1016/j.patcog.2018.10.026
19. International Institute for Population Sciences (IIPS). National Family Health Survey (NFHS 5) 2019-20: India Report. IIPS; 2021. Available from: https://www.dhsprogram.com/pubs/pdf/FR375/FR375.pdf [Accessed on April 05, 2025]
20. Eaton J, Masquelier B. demogsurv: Demographic analysis of DHS and other household surveys [R package]. Version 0.2.6. 2025 [cited 2025 Jul 16]. Available from: https://github.com/mrc-ide/demogsurv/tree/0389352e6cdd366f9b1324a0ffe837081c587d86
21. Bongaarts J. Human population growth and the demographic transition. Philos Trans R Soc Lond B Biol Sci. 2009; 364(1532):2985-2990. DOI: https://doi.org/10.1098/rstb.2009.0137 PMid:19770150 PMCid:PMC2781829
22. Coale AJ, Hill AG, Trussell TJ. A new method of estimating standard fertility measures from incomplete data. Population Index. 1975;41(2):182-210. DOI: https://doi.org/10.2307/2734617
23. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol 1. Berkeley (CA): Univ Calif Press; 1967. p. 281-297. Available from: http://projecteuclid.org/euclid.bsmsp/12005 12992
24. Hartigan JA, Wong MA. Algorithm AS 136: a K means clustering algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100-108. DOI: https://doi.org/10.2307/2346830
25. Onumanyi AJ, Molokomme DN, Isaac SJ, et al. AutoElbow: An automatic elbow detection method for estimating the number of clusters in a dataset. Appl Sci. 2022;12(15):7515. DOI: https://doi.org/10.3390/app12157515
26. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20(1):53-65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7
27. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001;63(2):411-426. DOI: https://doi.org/10.1111/1467-9868.00293
28. Krzanowski W, Lai Y. A criterion for determining the number of groups in a data set using sum of squares clustering. Biometrics. 1988;44(1):23-34. DOI: https://doi.org/10.2307/2531893
29. Patil C, Baidari I. Estimating the optimal number of clusters k in a dataset using data depth. Data Sci Eng. 2019;4(2):132-140. DOI: https://doi.org/10.1007/s41019-019-0091-y
30. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. J Stat Softw. 2014;61(6):1-36. DOI: https://doi.org/10.18637/jss.v061.i06
31. Hopkins B, Skellam JG. A new method for determining the type of distribution of plant individuals. Ann Bot. 1954;18(2):213-228. DOI: https://doi.org/10.1093/oxfordjournals.aob.a083391
32. Lawson RG, Jurs PC. New index for clustering tendency and its application to chemical problems. J Chem Inf Comput Sci. 1990;30(1):36-41. DOI: https://doi.org/10.1021/ci00065a010