Women's Education in Ethiopia: Distribution, Spatial Clustering, and Structural Implications*

Abstract: *Women's education (WE) is a foundational determinant of population health, economic participation, and intergenerational wellbeing. This study examines the distribution and spatial structure of women's educational attainment in Ethiopia using data from the 2016 Ethiopia Demographic and Health Survey (EDHS). The analysis integrates survey-weighted national and regional estimates with cluster-level spatial analysis and hot-spot detection using the Getis-Ord Gi statistic.*

Nationally, educational attainment reflects a broad base of low education, with substantial proportions of women lacking formal schooling and limited progression to secondary and higher levels. Regional analysis reveals marked geographic disparities, with the highest levels of deprivation in Somali and Afar and the lowest in Addis Ababa. However, cluster-level analysis demonstrates extreme local heterogeneity, with the percentage of women with no education ranging from 0% to 100% across 643 sampling clusters.

Spatial hotspot analysis identifies statistically significant clustering of educational deprivation in eastern and northeastern Ethiopia, particularly in Somali and Afar, alongside cold spots of high attainment in Addis Ababa and pockets in Tigray and Gambela.

The findings show that women's education in Ethiopia is not only unevenly distributed but spatially organized into localized systems of advantage and deprivation. This spatial structuring does not align neatly with conventional urban–rural or livelihood-based classifications, suggesting the influence of complex, place-specific factors. The results underscore the need for geographically targeted interventions and demonstrate the value of integrating spatial analysis into demographic research and policy planning.

*By Aynalem Adugna, Research Scientist, State of California

1. Introduction

Women's education is widely recognized as one of the most consequential structural determinants of population health, economic development, and intergenerational wellbeing. A large body of research demonstrates that increases in women's educational attainment are associated with improved maternal and child health outcomes, reduced fertility, enhanced labor force participation, and greater household-level decision-making power [1][2][3]. These effects operate through multiple pathways, including increased health knowledge, improved access to information, delayed marriage and childbearing, and strengthened agency in household and community contexts.

Unlike short-term behavioral indicators such as exclusive breastfeeding (EBF), which can change over relatively short time horizons in response to programmatic interventions or shifts in norms, education reflects long-run structural processes. These processes include the expansion

and quality of schooling systems, geographic accessibility of educational infrastructure, household economic capacity, and broader institutional investments. As a result, educational attainment evolves gradually and exhibits strong persistence over time and space [4][5]. This temporal and spatial persistence makes women's education a particularly important indicator for understanding underlying structural inequalities.

In Ethiopia, substantial progress has been made in expanding access to primary education over the past two decades, particularly following the implementation of the Education Sector Development Programs (ESDP) and broader pro-poor development strategies [6]. Enrollment rates have increased significantly, and gender gaps in primary education have narrowed. However, these gains have not translated uniformly into higher levels of completed education, especially at secondary and tertiary levels. Moreover, progress has been uneven across regions and population groups, with persistent disparities linked to geography, livelihood systems, and socioeconomic conditions [7][8].

National averages of women's education, while useful for tracking overall progress, can obscure these disparities. Aggregated statistics may suggest gradual improvement, yet conceal substantial variation across regions and, more importantly, across localities within regions. Recent studies emphasize that educational outcomes in Ethiopia are shaped not only by broad regional differences but also by highly localized conditions, including infrastructure access, settlement patterns, cultural practices, and service delivery capacity [9][10]. This implies that understanding women's education requires moving beyond national and regional summaries to examine finer spatial scales.

This section analyzes women's education in Ethiopia using the Ethiopia Demographic and Health Survey (EDHS) 2016, with a specific focus on spatial structure and clustering. The analysis proceeds in three stages: (1) national weighted distribution, (2) regional variation, and (3) cluster-level spatial analysis using hot-spot methods. By identifying statistically significant clusters of high and low educational attainment, the analysis aims to reveal the geographic organization of educational inequality.

The central argument advanced in this section is that women's education in Ethiopia is not merely unevenly distributed but is spatially structured and locally clustered, reflecting underlying systems of advantage and deprivation. This perspective shifts the focus from aggregate disparities to geographically embedded patterns, providing a more precise foundation for targeted policy interventions and for interpreting the relationship between structural determinants, such as education, and behavioral outcomes, such as exclusive breastfeeding (EBF).

2. Data and Methods

2.1 Data Source

This analysis uses data from the Ethiopia Demographic and Health Survey (EDHS) 2016 Individual Recode (IR) file, which contains individual-level information on women aged 15–49. The IR file is the appropriate data source for examining women's education because educational attainment is measured at the respondent (woman) level rather than at the household or child level.

The EDHS 2016 employed a two-stage stratified sampling design. In the first stage, enumeration areas (clusters) were selected using probability proportional to size. In the second stage, households were systematically sampled within each selected cluster. All eligible women in sampled households were interviewed. The survey is designed to produce estimates that are representative at both the national and regional levels [11][12].

The dataset used in this analysis includes:

- 15,683 women with valid interview records
- 643 sampling clusters, representing geographically distinct enumeration areas
- Stratification by region and urban/rural residence

The use of the IR file ensures that education is measured directly at the individual level, avoiding aggregation biases that would arise from using household-level files.

2.2 Variable Construction

Women's Education Variable

The primary variable of interest is:

- v106: highest educational level attained

This variable is defined in the DHS as a categorical indicator with the following values:

Code	Category
0	No education/preschool
1	Primary
2	Secondary
3	Higher

8	Don’t know

This recoding ensures that only valid educational attainment categories are included in the analysis. Observations coded as “don’t know” or missing are excluded to maintain interpretability and internal consistency.

Derived Indicator: No Education

To facilitate interpretation and spatial analysis, a binary indicator was constructed:
edu == 0
This identifies women with no formal education, which is used to compute cluster-level proportions.

Cluster-Level Aggregation
Cluster-level estimates were computed as:
pct_no_edu = mean(edu == 0) * 100
This produces, for each cluster:
The percentage of women with no education
Cluster-level aggregation serves two purposes:

i. It reduces individual-level variability to a spatially interpretable unit
ii. It enables linkage with DHS GPS coordinates for spatial analysis

This approach is consistent with prior DHS-based spatial analyses examining localized patterns in demographic and health indicators [13].

2.3 Survey Design

DHS data are not derived from simple random sampling. Instead, they reflect a complex survey design involving stratification, clustering, and unequal sampling probabilities. Failure to account for this design leads to biased estimates and incorrect standard errors.

To address this, survey-weighted estimation was implemented using the survey package in R:

- v005 (normalized as weight) represents the sampling weight

Sampling weights were normalized by dividing by 1,000,000, following DHS conventions:

weight = v005 / 1000000

Survey-weighted estimates were used for:

- National distributions
- Regional comparisons

Cluster-level calculations, by contrast, are unweighted within clusters, as they represent local proportions rather than population-level estimates.

This distinction is important:

- Weighted estimates → represent population-level distributions
- Cluster-level estimates → represent local spatial conditions

The combined use of both approaches allows for a comprehensive understanding of both aggregate and localized patterns [12][14].

2.4 Spatial Hot-Spot Analysis

Conceptual Framework

To assess spatial clustering, the analysis employs the Getis-Ord Gi* statistic, a widely used method for identifying statistically significant clusters of high or low values in spatial data. Unlike simple mapping of values, the Gi* statistic assesses whether observed clustering exceeds what would be expected under spatial randomness [15][16].

This approach is particularly appropriate for DHS data because:

- Clusters are irregularly spaced
- The objective is to detect spatial concentration, not smooth surfaces
- Policy relevance lies in identifying priority zones, not interpolated values

Implementation

Cluster-level data were joined to DHS GPS coordinates using the cluster identifier (v001). The resulting spatial dataset was analyzed in ArcGIS Pro using ArcPy.

Parameter Selection

K Nearest Neighbors (k = 8)
Defines spatial relationships based on each cluster's eight nearest neighbors. This is appropriate for DHS data because cluster density varies across regions, and fixed distance thresholds can produce unstable results.

Row Standardization
Ensures comparability across clusters by standardizing weights within each neighborhood.

Euclidean Distance
Used to measure straight-line distance between clusters. Given the national scale of the analysis, this is sufficient for identifying broad spatial patterns.

False Discovery Rate (FDR) Correction
FDR correction adjusts for multiple hypothesis testing, reducing the likelihood of false positives. This is particularly important when analyzing a large number of spatial units [16].

Interpretation of Outputs

The Gi* statistic produces:

- Hot spots: clusters with high values surrounded by high values
- Cold spots: clusters with low values surrounded by low values
- Non-significant areas: no strong spatial clustering

Significance is typically reported at:

- 90% confidence
- 95% confidence
- 99% confidence

These outputs allow for the identification of statistically meaningful spatial clusters, rather than relying on visual inspection alone.

Justification for Method Choice

Hot-spot analysis was selected over interpolation methods (e.g., Inverse Distance Weighting - IDW or kriging) because:

- DHS clusters do not form a continuous surface
- Interpolation may introduce artificial smoothing
- The research objective is cluster detection, not prediction

This aligns the method with both the structure of the data and the substantive research question.

3. Results

3.1 National Distribution

The national distribution of women's education in Ethiopia reveals a broad base of low educational attainment with limited progression to higher levels. Weighted estimates from the EDHS 2016 indicate that a substantial proportion of women of reproductive age have no formal education, while primary education constitutes the largest share among those with schooling. Secondary and higher education levels remain comparatively limited.

This distribution reflects a system that has achieved considerable expansion in access—particularly at the primary level—without a commensurate increase in completion and progression to higher levels of education. Such a pattern is consistent with national education-sector analyses, which document rapid gains in enrollment alongside persistent challenges in retention and transitions beyond primary schooling [17][18].

From an analytical standpoint, the national distribution provides an important baseline. However, as shown in subsequent sections, it conceals substantial spatial heterogeneity. The relatively moderate national average of educational attainment masks localized contexts in which educational deprivation is either near-universal or nearly absent.

3.2 Regional Distribution

Regional analysis reveals pronounced geographic variation in women's educational attainment. The proportion of women with no education varies substantially across Ethiopia's administrative regions, with a clear gradient emerging between eastern, central, and urban areas.

Table 1. Regional cluster means

Region	Mean % No Education
Somali	76.3
Afar	73.0
Amhara	55.5
Oromia	51.1
Benishangul	50.6
SNNPR	46.0
Tigray	42.9
Harari	33.5
Gambela	33.0
Dire Dawa	32.0
Addis Ababa	8.5

Regions such as Somali and Afar exhibit the highest levels of educational deprivation, with large shares of women lacking formal education. In contrast, Addis Ababa shows markedly lower levels of no education, reflecting the advantages associated with urban infrastructure, service

availability, and economic opportunities. Regions such as Amhara and Oromia occupy intermediate positions, with moderate levels of educational attainment.

Importantly, the observed regional pattern does not conform strictly to a simple urban–rural dichotomy. While urban areas generally demonstrate higher educational attainment, there is considerable variation within both rural and semi-urban regions. For example, some predominantly rural regions display moderate educational outcomes, while certain smaller urban or peri-urban areas exhibit lower-than-expected attainment.

This suggests that regional differences are shaped not only by urbanization but also by a combination of factors, including historical investment in education, geographic accessibility, livelihood systems, and institutional capacity. Similar patterns of uneven regional development have been documented in Ethiopia's education sector, where disparities reflect both structural and geographic constraints [18][19].

3.3 Cluster-Level Distribution

Cluster-level analysis reveals the most striking feature of women's education in Ethiopia: extreme spatial heterogeneity at the local level.

A total of 643 sampling clusters were analyzed. The percentage of women with no education within clusters exhibits the following distribution:

- Mean: 47.0%
- Median: 48.4%
- Minimum: 0%
- Maximum: 100%

This full range—from clusters with no women lacking education to clusters where all women lack education—demonstrates that educational attainment is not evenly distributed even within regions. Rather, it is highly localized.

This indicates that half of all clusters fall within a very wide band of educational attainments, reflecting substantial inequality across localities.

Cluster-level categorization reinforces this pattern:

- A fifth to one-quarter of sampling clusters fall into "very low" deprivation (<20% of women with no education), indicating relatively high educational attainment
- Just under 15% fall into "very high" deprivation (≥80% of women with no education), indicating near-universal lack of education in communities represented by over 90 sampling clusters.

Table 2. Distribution of Sampling Clusters by Educational Deprivation

Row Labels	Count of No-Educ	Percent No-Educ
Very Low (<20%)	152	23.6
Low (20-39%)	105	16.3
Moderate (40-59%)	156	24.3
High (60-79%)	137	21.3
High (80-100%)	93	14.5
Grand Total	**643**	**100.0**

The interquartile distribution further highlights this variability:

- 25th percentile: ~22%
- 75th percentile: ~70%

These findings demonstrate that Ethiopia's educational landscape is best understood as a mosaic of localized systems, rather than a smooth regional gradient. Similar cluster-level disparities have been noted in spatial analyses of DHS data, where local conditions often diverge significantly from regional averages [20].

3.4 Spatial Hot-Spot Results

The hot-spot analysis provides a statistically rigorous assessment of whether high or low levels of educational deprivation are spatially clustered (see Map 1).

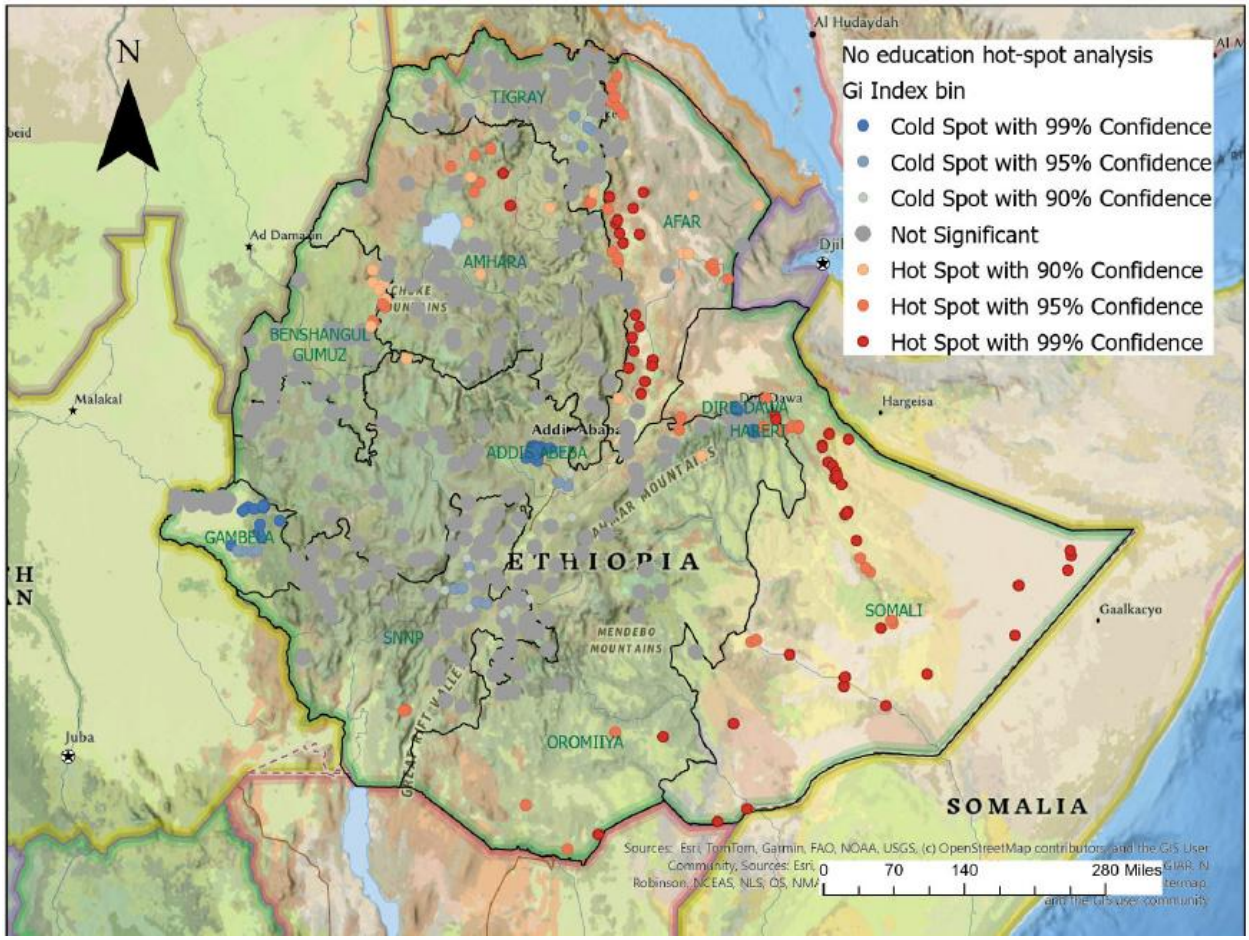
Hot spots (high %no education)

The analysis identifies a pronounced and contiguous cluster of high values in eastern and northeastern Ethiopia, particularly:

- Somali region (strongest and most extensive hot spot)
- Afar region
- Portions of the eastern corridor near Harari and Dire Dawa (Map 1)

These areas exhibit statistically significant clustering at high confidence levels, indicating that high levels of educational deprivation are not randomly distributed but spatially concentrated.

Map 1: Low-education Hot and Cold Spots: Results of a Getis-Ord Gi Analysis



Cold spots (low no education)

Statistically significant clusters of low values are observed in:

- Addis Ababa and surrounding central areas
- A few locations in Tigray, Harari and Gambela Kilils, as well as Dire Dawa City administration.

These cold spots represent localized concentrations of relatively high educational attainment.

Non-significant areas

A substantial portion of the country is classified as non-significant. This does not imply uniformity; rather, it indicates that in these areas, high and low values do not form sufficiently strong local clusters to be statistically distinguished from spatial randomness.

Interpretation

The hot-spot analysis confirms that:

Women's educational deprivation in Ethiopia is spatially clustered rather than randomly distributed.

The eastern regions form a coherent spatial system of disadvantage, while urban and selected peripheral regions form clusters of relative advantage.

3.5 Robustness to False Discovery Rate (FDR) Adjustment

To assess the robustness of the spatial clustering results, the hot-spot analysis was re-estimated using False Discovery Rate (FDR) correction, which adjusts for multiple hypothesis testing.

The application of FDR did not materially alter the spatial pattern of results:

- Core hot spots in Somali and Afar remained statistically significant
- Core cold spots in Addis Ababa, locations in Tigray, Harar and Gambela Kilils, as well as Dire Dawa City administration remained intact
- Only minor changes were observed in marginal clusters

Implications

The stability of the results under FDR adjustment indicates that:

The observed spatial clustering of women's educational deprivation is **robust and not driven by multiple testing artifacts**.

This strengthens the validity of the findings and supports the conclusion that the identified clusters reflect underlying structural conditions rather than statistical noise.

Summary of Findings

Across all levels of analysis, a consistent picture emerges:

1. National averages indicate moderate overall attainment
2. Regional analysis reveals broad geographic gradients

3. Cluster-level analysis shows extreme local heterogeneity
4. Hot-spot analysis confirms statistically significant spatial clustering
5. FDR adjustment demonstrates robustness of these patterns

Together, these results establish that women's education in Ethiopia is not only unevenly distributed but spatially organized into clusters of advantage and deprivation.

Below is a **fully expanded, synthesis-level Section 4 (Discussion)**, aligned with your results, your earlier EBF findings, and your methodological stance. It deepens interpretation, integrates spatial and substantive insights, and incorporates your critique of conventional regression approaches.

4. Discussion

4.1 Women's Education as a Structural System

The findings presented in this study reinforce the conceptualization of women's education as a slow-moving structural system rather than a proximate or short-term determinant of behavior. Unlike indicators such as EBF, which can respond relatively quickly to programmatic interventions, education reflects cumulative processes that unfold over long-time horizons. These processes include investments in schooling infrastructure, teacher availability, household economic capacity, and broader institutional development [8][10].

The national distribution of women's education in Ethiopia—characterized by a large share of women with no formal education and limited progression to secondary and higher levels—reflects both historical constraints and ongoing structural challenges. While the expansion of primary education has increased access, it has not yet fully translated into widespread educational attainment. This pattern is consistent with analyses of education systems that describe Ethiopia's trajectory as one of expansion without full transition, in which enrollment gains outpace completion and progression [17][19].

From an analytical perspective, this structural nature implies that women's education should not be interpreted as an immediately responsive explanatory variable in behavioral models. Rather, it functions as a background condition, shaping the context within which behaviors occur but not necessarily determining them directly.

4.2 Spatial Clustering as the Central Finding

The most important empirical contribution of this analysis lies in identifying statistically significant spatial clustering of women's educational deprivation. While national and regional analyses reveal gradients, the hot-spot analysis demonstrates that these gradients are organized into coherent geographic clusters.

The eastern regions of Ethiopia—particularly Somali and Afar—emerge as a contiguous spatial system of high educational deprivation. These areas are not merely characterized by higher averages; rather, they exhibit statistically significant concentrations of clusters with high proportions of women lacking education. The persistence of these clusters under FDR adjustment further confirms that they reflect underlying structural realities rather than random variation.

Conversely, cold spots of low educational deprivation are concentrated in Addis Ababa, locations in Tigray, Harari and Gambela Kilils, as well as Dire Dawa City administration, indicating localized systems of relative advantage. The coexistence of these hot and cold spots suggests that Ethiopia's educational landscape is best understood as a spatially segmented system, where both deprivation and advantage are geographically organized. This finding shifts the analytical focus from "how much inequality exists" to "how inequality is spatially structured," with important implications for both interpretation and policy design.

4.3 Beyond Regional Averages

A key insight from this analysis is that regional averages obscure substantial heterogeneity within regions. While regions such as Somali and Afar exhibit high average levels of educational deprivation, cluster-level analysis shows that even within these regions, there is variation. Similarly, regions with moderate averages, such as Oromia and SNNPR, contain clusters spanning a wide range of educational attainment.

This pattern underscores the limitations of relying solely on administrative boundaries for analysis and intervention. Regions are useful for governance and reporting, but they do not necessarily correspond to homogeneous social or developmental contexts. Instead, educational outcomes are shaped by localized conditions, including access to schools, transportation infrastructure, settlement patterns, and community-level socioeconomic dynamics.

The presence of both high- and low-performing clusters within the same region suggests that interventions based solely on regional targeting may be insufficiently precise. More granular approaches that identify and address local pockets of deprivation are likely to be more effective.

4.4 Localized Systems of Advantage and Deprivation

The cluster-level and hot-spot results indicate that women's education in Ethiopia is organized into localized systems of advantage and deprivation. These systems are likely driven by a combination of community-level factors.

In areas identified as hot spots—particularly Somali and Afar—several structural characteristics may contribute to low educational attainment:

- Geographic isolation and limited access to educational infrastructure
- Livelihood systems that may constrain formal schooling participation
- Historical underinvestment in education services
- Lower availability of secondary and post-primary schooling opportunities

At the same time, the presence of cold spots in regions such as Addis Ababa, locations in Tigray, Harari and Gambela Kilils, as well as Dire Dawa City administration, suggests that localized conditions can support higher levels of educational attainment even within a broader national context of constraint.

Importantly, these localized systems are not static. They represent outcomes of cumulative processes and may evolve over time with changes in infrastructure, policy, and economic conditions. However, their persistence across clusters indicates that they are deeply embedded and not easily altered through short-term interventions.

4.5 Relationship to Behavioral Outcomes (WE vs EBF)

A central analytical implication of this study is the distinction between **structural determinants** (such as women's education) and **behavioral outcomes** (such as exclusive breastfeeding). While education is often included as an explanatory variable in regression models of health behavior, the findings here suggest that its role may be more complex.

The EBF analysis conducted alongside this study showed that:

- EBF declines sharply with infant age
- Regional variation exists but is not always aligned with socioeconomic gradients
- Education and wealth exhibit weak or inconsistent associations with EBF

In contrast, women's education exhibits:

- Strong spatial clustering
- Stability across analytic specifications
- Robustness to statistical adjustments (e.g., FDR)

This contrast suggests that:

Education functions as a **context-setting variable**, while behaviors such as EBF are shaped by more immediate, situational factors.

4.6 Methodological Implications: Limits of Conventional Regression

These findings also have implications for the interpretation of regression-based studies of health and education outcomes. Many studies—including those using DHS data—employ multivariable regression models to identify "determinants" of behaviors or outcomes. However, such models often assume that explanatory variables operate independently and uniformly across contexts.

In practice, variables such as education, wealth, and media exposure are highly correlated and often reflect overlapping dimensions of socioeconomic status. Failure to adequately account for these interrelationships—through tests of collinearity or alternative modeling approaches—can lead to unstable estimates and potentially misleading conclusions.

Furthermore, regression models typically do not account for spatial dependence, meaning that observations located near each other may be more similar than those farther apart. When spatial clustering is present—as demonstrated in this analysis—standard regression assumptions are violated, and estimated relationships may be biased or misinterpreted.

The present analysis addresses these limitations by:

- Examining distributions directly (rather than relying solely on model coefficients)
- Incorporating spatial analysis to identify clustering
- Distinguishing between structural and behavioral variables

This approach provides a more nuanced understanding of how education operates within a broader system of inequality.

4.7 Policy Implications

The findings of this study have several important implications for policy and program design.

First, the identification of spatial hot spots suggests that interventions should prioritize geographically concentrated areas of deprivation, particularly in Somali and Afar regions. Broad national or regional strategies may be insufficient to address localized challenges.

Second, the existence of cold spots indicates that successful models of educational attainment exist within the country. Understanding the conditions that support higher attainment in these areas may provide valuable lessons for replication elsewhere.

Third, the weak alignment between education and behavioral outcomes such as EBF suggests that improving education alone may not be sufficient to change specific health behaviors in the short term. Complementary interventions addressing cultural norms, service access, and immediate constraints are likely to be necessary.

Finally, the results underscore the importance of integrating spatial analysis into routine monitoring and evaluation. Identifying and tracking spatial patterns can enhance the targeting and effectiveness of interventions.

Below is your final, fully developed Section 5 (Conclusion)—expanded, nuanced, and tightly aligned with your empirical results, your hot-spot findings, and your earlier corrections regarding oversimplified narratives (e.g., urban–rural or pastoralist framing).

5. Conclusion

This study set out to examine women's education in Ethiopia through a multi-level lens—national, regional, and cluster-level—while incorporating spatial statistical methods to identify patterns of concentration and dispersion. The results collectively demonstrate that women's education in Ethiopia is not only unevenly distributed but is **spatially structured and locally clustered in statistically meaningful ways**.

At the national level, the distribution of educational attainment reflects a system in transition. While access to primary education has expanded significantly, a large proportion of women remain without formal education, and progression to secondary and higher levels remains limited. This pattern is consistent with a broader trajectory of educational expansion that has not yet translated into widespread attainment.

However, the most important findings emerge at finer levels of analysis. Regional variation reveals broad geographic gradients, but cluster-level analysis exposes the full extent of inequality. The observed range—from clusters where no women lack education to clusters where all women lack education—indicates that educational attainment is highly localized. This degree of heterogeneity cannot be captured by national or regional summaries alone.

The spatial hot-spot analysis provides a critical layer of inference by demonstrating that these differences are not randomly distributed. Instead, they form **coherent geographic clusters of advantage and deprivation**. Eastern Ethiopia—particularly Somali and Afar—emerges as a statistically significant and contiguous zone of high educational deprivation, while areas such as Addis Ababa, locations in Tigray, Harari and Gambela Kilils, as well as Dire Dawa City administration form clusters of relatively high attainment. The persistence of these patterns under False Discovery Rate correction confirms that they represent robust spatial structures rather than artifacts of statistical testing.

A key contribution of this study is the clarification that these spatial patterns do not align neatly with simplified categorizations such as urban versus rural or pastoralist versus non-pastoralist systems. For example, regions with similar livelihood structures can exhibit different levels of educational attainment, and areas with urban characteristics do not uniformly display high

education levels. This suggests that women's education is shaped by a **complex interaction of historical investment, institutional capacity, geographic accessibility, and localized socioeconomic conditions**, rather than by any single explanatory dimension.

The findings also have important implications for how education is interpreted in relation to behavioral outcomes. In contrast to exclusive breastfeeding—where patterns are dynamic and influenced by immediate contextual factors—women's education exhibits stability and strong spatial clustering. This reinforces the view that education operates as a **structural, context-setting variable** rather than a direct, short-term determinant of behavior. As such, its influence is likely to be indirect, long-term, and mediated through multiple pathways.

Methodologically, the study highlights the value of combining descriptive, spatial, and inferential approaches. By moving beyond regression-based frameworks and incorporating spatial clustering analysis, the study avoids assumptions of independence and uniformity that may not hold in geographically structured data. This approach provides a more accurate representation of how inequality is organized in space and how it manifests at the local level.

From a policy perspective, the results underscore the importance of **geographically targeted interventions**. National and regional strategies, while necessary, are insufficient on their own to address deeply localized disparities. The identification of statistically significant hot spots offers a practical basis for prioritizing areas of greatest need, particularly in Somali and Afar. At the same time, the presence of cold spots suggests that localized success stories exist and can inform context-specific strategies for improving educational outcomes.

In conclusion, women's education in Ethiopia is best understood not as a uniform national characteristic or even a simple regional gradient, but as a **spatially organized system of localized advantage and deprivation**. Addressing educational inequality therefore requires approaches that recognize and respond to this spatial complexity. Future research and policy efforts should continue to integrate spatial analysis with traditional demographic methods, enabling more precise identification of needs and more effective allocation of resources.

Appendix: Core R Workflow (Teaching Version with Line-by-Line Notes)

A. What this script does

This script does five main things:

1. Opens the Ethiopia DHS 2016 women's data file
2. Keeps only the variables needed for women's education analysis
3. Creates a clean education variable
4. Produces national and regional weighted results
5. Produces cluster-level values for mapping and hot-spot analysis

B. Very important symbol: `%>%`

Before the code, here is the most important symbol to understand:

`%>%`

This is called the **pipe**.

How to read it

You can read `%>%` as:

- "then"
- "and next do this"
- "take the result from the previous line and pass it to the next line"

Example

```
data %>%
  filter(x == 1) %>%
  select(y, z)
```

This means:

1. start with `data`
2. then keep only rows where `x == 1`
3. then keep only columns `y` and `z`

So `%>%` helps R code read like a sequence of steps rather than nested commands.

C. Load packages

```
# Load the haven package.
# "library()" tells R to make a package available for use.
# The haven package lets R read Stata files such as DHS.DTA
files.
library(haven)

# Load dplyr.
# dplyr is used for data management:
# selecting columns, creating variables, filtering rows,
grouping, and summarizing.
library(dplyr)

# Load survey.
# survey is needed because DHS data come from a complex sample
design,
# not from a simple random sample.
library(survey)

# Load forcats.
# forcats helps with factor variables (categorical variables
with labels).
library(forcats)
```

D. Read the DHS data file

```
# Create an object called path.
# Think of "path" as the address of the data file on the
computer.
path <-
"C:/Users/aynal/Documents/DHS2016/Data/ETIR71DT/ETIR71FL.DTA"

# Read the DHS Stata file into R.
# read_dta() comes from the haven package.
# The result is stored in an object called ir.
# "ir" stands for Individual Recode.
ir <- read_dta(path)
```

How to read this

- `<-` means "put the result into"
- So:
    - `path <- ...` means "store this file location in `path`"
    - `ir <- read_dta(path)` means "read the file and store it in `ir`"

E. Keep only the variables needed

```
# Create a smaller working dataset called we vars (we stands for
women's education).
# This keeps only the variables needed for the women's education
analysis.
we_vars <- ir %>%

  # select() keeps only the columns listed inside it.
  # v001 = cluster ID
  # v021 = primary sampling unit (PSU)
  # v022 = sampling strata
  # v024 = region
  # v005 = sampling weight
  # v106 = highest education level
  select(v001, v021, v022, v024, v005, v106) %>%

  # mutate() creates new variables or changes existing ones.
  mutate(

    # DHS weights are stored as large integers.
    # Divide by 1,000,000 to convert to a proper weight.
    weight = v005 / 1000000,

    # Create a simpler education variable called edu.
    # case_when() means:
    # "if this condition is true, assign this value"
    edu = case_when(
      v106 == 0 ~ 0,   # no education
      v106 == 1 ~ 1,   # primary
      v106 == 2 ~ 2,   # secondary
      v106 == 3 ~ 3,   # higher

      # If the value is something else, such as "don't know",
      # store it as missing (NA).
      TRUE ~ NA_real_
    ),

    # Convert region codes into readable labels.
    # For example, 1 becomes Tigray, 2 becomes Afar, etc.
    region = as_factor(v024)
  )
```

How to read `case_when()`

This line:

```
v106 == 0 ~ 0
```

means:

- if `v106` equals `0`
- then assign the value `0` to the new variable `edu`

The last line:

```
TRUE ~ NA_real_
```

means:

- if none of the earlier conditions matched
- assign missing numeric value (`NA`)

F. Look at the education variable

```
# table() counts how many times each value appears.
# useNA = "ifany" tells R to also show missing values if they
exist.
table(we_vars$edu, useNA = "ifany")
```

How to read `we_vars$edu`

The `$` sign means:

- go into the object `we_vars`
- and use the variable `edu`

So `we_vars$edu` means:

the `edu` column inside the `we_vars` dataset

G. Set up the DHS survey design

```
# Create a survey design object called design_we.
# This tells R how the DHS sample was selected.
design_we <- svydesign(

  # id = cluster or primary sampling unit
  id = ~v021,
```

```
  # strata = sampling strata
  strata = ~v022,

  # weights = sample weights
  weights = ~weight,

  # data = the dataset being used
  data = we_vars,

  # nest = TRUE helps R handle the sample design correctly
  nest = TRUE
)
```

How to read `~v021`

The tilde ~ is used in formulas in R.

So:

```
id = ~v021
```

means:

use variable `v021` as the cluster ID in the survey design

H. National weighted distribution of women's education

```
# svymean() calculates weighted means or proportions using
survey design.
# factor(edu) tells R that education is a category, not a
continuous number.
we_national <- svymean(~factor(edu), design_we, na.rm = TRUE)

# Print the result to the screen.
print(we_national)
```

How to read this

- `svymean()` = survey-weighted mean/proportion
- `~factor(edu)` = treat `edu` as categories
- `design_we` = use the DHS survey design
- `na.rm = TRUE` = ignore missing values

I. Turn national results into percentages

```
# Convert the national result into a regular data frame
# so that it is easier to work with.
we_national_df <- as.data.frame(we_national)

# Multiply by 100 to turn proportions into percentages.
we_national_df <- we_national_df %>%
  mutate(
    percent = `factor(edu)` * 100
  )

# Print the table.
print(we_national_df)
```

Important note

Depending on your R version, the column name may not appear exactly as `factor(edu)`. If this line gives an error, first run:

```
names(we_national_df)
```

Then replace `factor(edu)` with the actual column name shown by R.

J. Regional weighted distribution

```
# svyby() means:
# calculate a survey-weighted result BY groups.
# Here, the groups are regions.
we_region <- svyby(

  # Outcome variable: education categories
  ~factor(edu),

  # Grouping variable: region
  ~region,

  # Survey design object
  design_we,

  # Function to apply
  svymean,
```

```
  # Remove missing values
  na.rm = TRUE,

  # Also calculate standard errors and confidence intervals
  vartype = c("se", "ci")
)

# Print the regional table.
print(we_region)
```

How to read `svyby()`

You can read it as:

"calculate survey-weighted percentages of education, by region"

K. Create the cluster-level dataset

```
# Build a new dataset called cluster_we.
# Each row will now represent one cluster, not one woman.
cluster_we <- we_vars %>%

  # Keep only women with valid education values.
  filter(!is.na(edu)) %>%

  # group_by(v001) means:
  # put all women from the same cluster together.
  group_by(v001) %>%

  # summarise() creates one summary row per cluster.
  summarise(

    # n = number of women in that cluster
    n = n(),

    # mean(edu == 0) calculates the proportion with no education
    # Multiply by 100 to express it as a percentage.
    pct_no_edu = mean(edu == 0) * 100,

    # .groups = "drop" tells R not to keep grouping afterward
    .groups = "drop"
  ) %>%

  # Join region back onto the cluster table.
  left_join(
    we_vars %>%
```

```
      select(v001, region) %>%
      distinct(),
    by = "v001"
  )
```

How to read `mean(edu == 0)`

This is one of the most useful tricks in R.

- `edu == 0` creates a TRUE/FALSE result
- TRUE means "this woman has no education"
- FALSE means "this woman does not"

In R:

- TRUE behaves like 1
- FALSE behaves like 0

So:

```
mean(edu == 0)
```

means:

the proportion of women with no education

Then multiplying by 100 gives percent.

L. Examine cluster-level results

```
# Show the first few rows
head(cluster_we)

# Summarise the distribution of cluster-level no education
summary(cluster_we$pct_no_edu)

# Count the number of clusters
nrow(cluster_we)
```

How to read these

- `head()` = show first few rows
- `summary()` = give minimum, quartiles, median, mean, maximum
- `nrow()` = count rows

Since each row is a cluster here:

`nrow(cluster_we)` = number of clusters

M. Regional summaries of cluster-level no education

```
# Create a regional summary table based on cluster values.
region_cluster_summary <- cluster_we %>%
  group_by(region) %>%
  summarise(
    clusters = n(),
    mean_pct_no_edu = mean(pct_no_edu),
    median_pct_no_edu = median(pct_no_edu),
    min_pct_no_edu = min(pct_no_edu),
    max_pct_no_edu = max(pct_no_edu),
    .groups = "drop"
  ) %>%

  # Sort from highest to lowest mean no education
  arrange(desc(mean_pct_no_edu))

# Print the summary
print(region_cluster_summary)
```

N. Create cluster categories for mapping

```
# Create labeled categories to simplify maps and tables.
cluster_we <- cluster_we %>%
  mutate(
    noedu_category = case_when(
      pct_no_edu < 20 ~ "Very Low (<20%)",
      pct_no_edu < 40 ~ "Low (20-39%)",
      pct_no_edu < 60 ~ "Moderate (40-59%)",
      pct_no_edu < 80 ~ "High (60-79%)",
      TRUE ~ "Very High (80-100%)"
    )
  )
```

Why use "-" instead of "–"?

The plain keyboard dash − avoids CSV/Excel encoding problems.

O. Count clusters in each category

```
# Count how many clusters fall in each education category band.
table(cluster_we$noedu_category)
```

P. Export for ArcGIS Pro

```
# Write the cluster dataset to a CSV file for GIS use.
write.csv(
  cluster_we,
  "cluster_we_noedu.csv",
  row.names = FALSE,
  fileEncoding = "UTF-8-BOM"
)
```

How to read this

- `write.csv()` = save data as CSV
- `row.names = FALSE` = do not add row numbers as a separate column
- `fileEncoding = "UTF-8-BOM"` = helps Excel open special characters correctly

Q. Simple quality checks

```
# Check for missing education values
sum(is.na(we_vars$edu))

# Check number of unique clusters
n_distinct(cluster_we$v001)

# Check the range of cluster-level percentages
range(cluster_we$pct_no_edu)
```

Appendix 2: ArcGIS Hot-Spot Analysis (Python /ArcPy)

Script

```
# Hot Spot Analysis (Getis-Ord Gi*)

arcpy.stats.HotSpots(
    Input_Feature_Class="MASTER_CLUSTER",   # DHS cluster point layer
    Input_Field="PCNT_NOED",               # % women with no education
    Conceptualization_of_Spatial_Relationships="K_NEAREST_NEIGHBORS",  #
define neighbors
    Distance_Method="EUCLIDEAN_DISTANCE",  # straight-line distance
    Standardization="ROW",                 # balance influence across
clusters
    Apply_False_Discovery_Rate__FDR__Correction="APPLY_FDR",  # adjust for
multiple testing
    number_of_neighbors=8                 # use 8 nearest clusters
)
```

References

[1] World Bank. (n.d.). *Education Overview*.
https://www.worldbank.org/en/topic/education/overview

[2] UNICEF. (Various years). *The State of the World's Children*.
https://www.unicef.org/reports/state-worlds-children

[3] UNESCO. (Annual). *Global Education Monitoring Report*.
https://www.unesco.org/gem-report/en

[4] Caldwell, J. C. (1979). Education as a factor in mortality decline: An examination of Nigerian data. *Population Studies*, 33(3), 395–413.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2861982/

[5] World Health Organization (WHO). (n.d.). *Social Determinants of Health*.
https://www.who.int/health-topics/social-determinants-of-health

[6] Government of Ethiopia & Ministry of Education. (Various phases). *Education Sector Development Program (ESDP)*.
https://www.globalpartnership.org/where-we-work/ethiopia

[7] UNICEF Ethiopia. (n.d.). *Education in Ethiopia: Overview*.
https://www.unicef.org/ethiopia/education

[8] World Bank. (Various years). *Ethiopia Country Overview*.
https://www.worldbank.org/en/country/ethiopia/overview

[9] Central Statistical Agency (CSA) [Ethiopia] and ICF. (2016). *Ethiopia Demographic and Health Survey 2016: Final Report*. Addis Ababa, Ethiopia and Rockville, Maryland, USA.
https://dhsprogram.com/publications/publication-FR328-DHS-Final-Reports.cfm

[10] United Nations Development Programme (UNDP). (Annual). *Human Development Report*.
https://hdr.undp.org

[11] The DHS Program (ICF). *DHS Methodology: Survey Design and Sampling*.
https://dhsprogram.com/methodology/Survey-Types/DHS-Methodology.cfm

[12] Rutstein, S. O., & Rojas, G. (2006). *Guide to DHS Statistics*. Calverton, MD: ORC Macro. https://dhsprogram.com/publications/publication-dhsg1-dhs-questionnaires-and-manuals.cfm

[13] Burgert-Brucker, C., et al. (2014). *Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys*. DHS Spatial Analysis Report No. 7. https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf

[14] Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied Survey Data Analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. https://www.routledge.com/Applied-Survey-Data-Analysis/Heeringa-West-Berglund/p/book/9781466565303

[15] ESRI. *Hot Spot Analysis (Getis-Ord Gi)* (Spatial Statistics)*. ArcGIS Pro Documentation. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/hot-spot-analysis.htm

[16] Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189–206. https://link.springer.com/article/10.1007/BF00162863

[17] World Bank. (n.d.). *Ethiopia General Education Quality Improvement Program and Education Sector Analysis*. https://www.worldbank.org/en/country/ethiopia/publication/ethiopia-general-education-quality-improvement-program

[18] UNICEF Ethiopia. (n.d.). *Education in Ethiopia: Overview*. https://www.unicef.org/ethiopia/education

[19] UNESCO. (Annual). *Global Education Monitoring Report*. https://www.unesco.org/gem-report/en

[20] The DHS Program. (Various years). *DHS Spatial Analysis Reports*. https://dhsprogram.com/publications/publication-search.cfm?type=5