Salud, Ciencia y Tecnología. 2025; 5:2180
doi: 10.56294/saludcyt20262180

ORIGINAL

Educative exclusion in Ecuador: identification of priority groups through strategic multivariate analysis for the formulation of public policies

Exclusión educativa en Ecuador: identificación de grupos prioritarios mediante análisis multivariante estratégico para la formulación de políticas públicas

Melissa Mercedes Idrovo-Hurel[1], Angel Antonio Morán-Herrera[1], Dennis Alfredo Peralta-Gamboa[1]

[1]Universidad Estatal de Milagro. Milagro EC0910150, Ecuador.

Cite as: Idrovo-Hurel MM, Morán-Herrera AA, Peralta-Gamboa DA. Educative exclusion in Ecuador: identification of priority groups through strategic multivariate analysis for the formulation of public policies. Salud, Ciencia y Tecnología. 2025; 5:2180. https://doi.org/10.56294/saludcyt20262180

Submitted: 03-07-2025 **Revised:** 05-09-2025 **Accepted:** 16-11-2025 **Published:** 17-11-2025

Editor: Prof. Dr. William Castillo-González

Corresponding author: Melissa Mercedes Idrovo-Hurel

ABSTRACT

This multivariate study analyzes educational attendance patterns in Ecuador, using data from the 2022 Population and Housing Census. The main purpose is to identify the age group with the highest incidence of educational exclusion, in order to guide targeted interventions. The research is classified as descriptive and exploratory, with a quantitative approach. The universe is composed of the population surveyed at the national level, and the unit of analysis corresponds to five-year age groups. The variables used were the total number of attendees, non-attendees, and total population, disaggregated by sex. The analysis integrates techniques such as graphical representation using principal components, hierarchization by Pareto analysis, and segmentation by hierarchical clusters, represented in heat maps. These tools made it possible to identify patterns of dependency and similarity between age groups, as well as to establish priorities for intervention. Among the main results, it is noteworthy that the 30-49 age groups account for more than 50 % of the population that does not attend educational institutions, suggesting a structural lag. The segmentation confirms the existence of different profiles between age groups. It is concluded that the proposed approach facilitates the identification of priority cohorts, providing useful evidence for the formulation of educational policies aimed at reintegrating adults into the education system.

Keywords: Educational Exclusion; Multivariate Analysis; Public Policies; Age Groups.

RESUMEN

Este estudio multivariante analiza los patrones de asistencia educativa en Ecuador, empleando datos del Censo de Población y Vivienda de 2022. El propósito principal es identificar el grupo etario con mayor incidencia de exclusión educativa, a fin de orientar intervenciones focalizadas. La investigación se clasifica como descriptiva y exploratoria, con enfoque cuantitativo. El universo está compuesto por la población censada a nivel nacional, y la unidad de análisis corresponde a los grupos quinquenales de edad. Se utilizaron como variables el número total de asistentes, no asistentes y población total, desagregadas por sexo. El análisis integra técnicas como representación gráfica mediante componentes principales, jerarquización por análisis de Pareto y segmentación por conglomerados jerárquicos, representados en mapas de calor. Estas herramientas permitieron identificar estructuras de dependencia y similitud entre grupos etarios, así como establecer prioridades de intervención. Entre los principales resultados, se destaca que los grupos entre 30 y 49 años concentran más del 50 % de la población que no asiste a centros educativos, lo que sugiere un rezago estructural. La segmentación confirma la existencia de perfiles diferenciados entre grupos de edad. Se concluye que el enfoque propuesto facilita la identificación de cohortes prioritarias, aportando evidencia útil para la formulación de políticas educativas dirigidas a la reintegración de personas adultas al sistema educativo.

Palabras clave: Exclusión Educativa; Análisis Multivariante; Políticas Públicas; Grupos Etarios.

INTRODUCTION

Education remains a fundamental right and a cornerstone of sustainable development. In recent decades, interest in improving the quality, equity, and inclusion of education has grown globally in response to the persistent inequalities that affect millions of people. Despite the efforts of Latin American countries to universalize school access, educational exclusion continues to affect historically marginalized groups.[1,2,3]

According to the Global Education Monitoring Report,[3] around 244 million children and young people worldwide do not attend school, and Latin America accounts for a significant percentage of this population. In Ecuador, data from the Population and Housing Census[1] show that 17 % of the population aged five and older does not attend any educational institution, with a particular concentration in the 30-49 age group, indicating a structural educational gap.

The concept of inclusive education, central to the Sustainable Development Goals (SDG 4), is not limited to access but encompasses equity in retention, meaningful participation, and completion of educational trajectories. However, various forms of exclusion threaten this ideal: early school dropout, lack of opportunities for adults, digital divides, economic precariousness, and structural discrimination.[1,4] Additionally, recent studies have demonstrated that geographic isolation and low connectivity have a detrimental impact on learning, particularly in rural areas.[5]

In Ecuador, although policies are in place to promote universal access, barriers to the full inclusion of adults persist, including a shortage of flexible programs, a lack of incentives for educational reintegration, and inadequate inter-institutional coordination. This situation highlights a historical debt to those who left the school system for economic, social, or personal reasons.[6,7,8,9,10,11]

Despite the importance of the problem, most studies have focused on childhood and adolescence, neglecting a specific analysis of educational exclusion among adults. In addition, few studies apply advanced statistical techniques to segment and rank the most affected groups.

The complexity of governing interdependent systems has also been addressed from transitional perspectives, as in [6] which can be extrapolated to the context of decentralized education policies. In the evaluation of public policies. From an institutional perspective,[7] review the link between the common good and the neo-Weberian state, highlighting its usefulness in designing inclusive policies. Various institutional configurations in higher education determine access according to specific educational regimes.[12,13,14,15,16,17]

This study seeks to fill that gap by generating recent evidence on the distribution and characteristics of educational exclusion based on age. The use of multivariate techniques allows for the identification of complex patterns that are not visible with traditional methods. By focusing on the adult population, this analysis provides key elements for rethinking educational policies from a logic of genuine inclusion and social sustainability.[18,19,20,21,22,23]

To analyze educational exclusion in Ecuador using multivariate statistical tools, to identify and prioritize the most affected age groups, and to contribute to the formulation of public policies aimed at the educational inclusion of adults.[24,25,26,27,28]

METHOD

Type of study

This work is part of a quantitative research design with a descriptive and exploratory approach. It is considered quantitative because it is based on the collection, processing, and analysis of large-scale numerical data from the national census. It is descriptive in that it aims to characterize the phenomenon of educational exclusion according to variables such as age, sex, and school attendance status, without intervening in or manipulating environmental conditions. It is exploratory because it seeks to identify latent patterns, underlying structures, and non-obvious relationships between variables using multivariate statistical techniques, which is essential for studies in areas with little recent empirical research, such as educational exclusion among the adult population in Ecuador.

This type of study is justified by the need to generate complex and segmented diagnoses that support evidence-based public decision-making. Unlike correlational or explanatory studies, the priority here is the structural detection of exclusion profiles and their subsequent visualization and prioritization for targeted educational policies.

Universe and sample

The study universe corresponds to the population residing in Ecuador registered in the VIII Population Census and VII Housing Census 2022, which is the most current and complete source of information on the demographic, educational, and social conditions of the country.

The sample analyzed was constructed from an aggregate census table published by the National Institute of Statistics and Censuses (INEC), which breaks down attendance at educational establishments by five-year age groups and sex. This classification results in a total of 27 observations (rows), corresponding to the following age groups:

- 0-4, 5-9, 10-14, 15-19, ..., up to 85 years and older.

Each of these groups represents a unit of analysis. Probabilistic sampling techniques were not applied, as complete census information was used, which lends robustness to the findings by not relying on statistical inferences but rather on direct analysis of the total population.

Variables analyzed

The analysis was based on four fundamental quantitative variables that reflect key dimensions of the educational phenomenon:

1. Total number of people attending an educational establishment, by age group.
2. Total number of people who do not attend any educational establishment, by age group.
3. Total male population, disaggregated by age.
4. Total female population, disaggregated by age.

These variables allow us to construct profiles of educational exclusion and inclusion, incorporating age and gender. Contextual or qualitative variables (such as level of education or reason for non-attendance) were excluded because they were not available in the selected base table.

For the purposes of multivariate statistical analysis, these variables were organized in an $X(n \times p)$ matrix, where $n = 27$ observations (age groups) and $p = 4$ variables. Prior to analysis, the data were standardized using Z scores to ensure comparability between variables with different scales.

Data source and processing

Data source

The information was obtained from the official website of the National Institute of Statistics and Census (INEC) of Ecuador, specifically from the national table of attendance at educational establishments by age group and sex, belonging to the VIII Population Census and VII Housing Census 2022. This source is official, public, and free of charge, so its use for academic purposes does not present any ethical or legal restrictions.

The table was downloaded in Excel format and then processed in a statistical environment for cleaning, organization, and transformation.

Statistical processing

The following phases were carried out:

Standardization of the data matrix

Each variable was transformed using Z scores, subtracting the mean and dividing by the standard deviation. This transformation allows all variables to contribute equally to the multivariate analyses, regardless of their original scale.

Principal Component Analysis (PCA)

PCA was applied to reduce the dimensionality of the dataset, capture the maximum variance explained in the first dimensions, and identify latent correlations between variables. The first two components explained more than 99 % of the total variance, allowing them to be visualized in a two-dimensional plane.

HJ-Biplot Representation

Once the data matrix was constructed and standardized, Principal Component Analysis (PCA) was applied as a basic technique to identify structural patterns in the information and reduce its dimensionality without substantial loss of information. The objective was to jointly represent the observations (age groups) and variables (attendance, non-attendance, population by sex) in a two-dimensional space, using an HJ-Biplot graph.

This representation, proposed by (26), helps visualize the geometric relationships between variables, the proximities between observations, and the dependency structures that are not easily perceived through univariate analyses. Its application enabled us to visually explore which age groups have similar profiles in terms of educational behavior and how these profiles relate to the included variables.

To implement this technique, the R programming language (version 4.4.2) was used, widely recognized in the scientific community for its power for statistical analysis and its ability to generate high-quality graphics. In particular, three specialized packages within R were used: FactoMineR, factoextra, and ggplot2.

The FactoMineR package is designed to perform multivariate statistical analyses in an accessible and reproducible manner. In this study, the PCA function was used to apply principal component analysis to the standardized matrix. This function automatically calculates the principal components, the coordinates of the observations, the contribution of each variable, and the proportion of variance explained, which facilitates the construction of geometric representations such as the HJ-Biplot.

Once the PCA model was obtained, the factoextra package, which acts as a graphical extension of FactoMineR, was used. This package enables the visualization of PCA results in a precise and customized manner using functions such as fviz_pca_biplot. In this study, this function was used to construct the HJ-Biplot graph, simultaneously representing the vectors of the variables and the points corresponding to the age groups. In addition, factoextra's ability to highlight contributions, adjust scales, group categories, and define thematic colors was used to facilitate the interpretation of the results.

The customization and visual enhancement of the graph were completed using the ggplot2 package, one of R's most versatile tools for generating statistical graphs. Through its syntax based on graphic layers (graphics grammar), ggplot2 allows users to modify aesthetic aspects such as color palettes, labels, text sizes, fonts, and the layout of graphic elements. In this study, the HJ-Biplot was refined to ensure a clear and aesthetically pleasing visualization suitable for interpretation in both academic and public decision-making contexts.

The result of this combination of tools was an HJ-Biplot graph that visually synthesizes the main findings of the PCA, facilitating the identification of age groups with similar educational profiles, proximity to certain variables, and the detection of structural patterns of academic exclusion. This visualization is a fundamental analytical resource in the multivariate approach adopted by the study.

Pareto analysis (80/20)

To rank age groups according to their level of educational exclusion, a Pareto analysis was applied to the variable "total number of people who do not attend educational establishments." This technique, based on the 80/20 principle originally formulated by [27] posits that in many social phenomena, a small proportion of causes generate most of the effects. In this case, the objective was to identify which age groups account for the highest cumulative volume of non-attendance at educational establishments, so that they can be prioritized in public intervention strategies.

The analysis consisted of ranking the age groups from highest to lowest according to the absolute number of people who do not attend, calculating the percentage that each group represents of the total, and generating a cumulative distribution that allows three categories to be established:

- Group A, representing the cumulative 80 % (high priority),
- Group B, between 80 % and 95 %,
- Group C, representing the remaining 5 % (low priority).

This process allowed the population to be segmented strategically, distinguishing the groups that have the most significant impact on national exclusion figures.

The R programming language (version 4.4.2) was also used to carry out this analysis. Although Pareto analysis can be performed with basic tools, this study employed a programmed and reproducible approach that enabled the technique to be seamlessly integrated into the overall statistical workflow of the survey. In particular, functions from the R base package and tools from the dplyr package, a library specializing in data manipulation, were used to group, sort, calculate proportions, and apply cumulative transformations in an efficient and readable manner.

The implementation in R consisted of sorting the variable of interest using the arrange function, calculating the percentages with the mutate function, and obtaining the cumulative sum with cumsum, which allowed the cutoff point between groups A, B, and C to be accurately identified. This classification was then incorporated as a categorical variable into the HJ-Biplot graph, assigning different colors to the groups according to their priority level.

In addition, to visualize the Pareto curve and graphically validate the cumulative distribution, the ggplot2 package was used to create a combined bar and line graph, which shows how a few age groups account for the majority of exclusion cases. This visualization is useful not only from an academic perspective but also as input for public policy formulation, as it clearly shows where institutional efforts should be concentrated.

This combined approach—hierarchical quantitative analysis and graphical visualization—allows us to rank the groups most affected by educational non-attendance with empirical evidence. Its integration with other multivariate techniques, such as PCA and hierarchical clustering, reinforces the consistency of the methodological approach adopted in the study.

https://doi.org/10.56294/saludcyt20262180 ISSN: 2796-9711

Hierarchical clustering analysis and heat map

From a historical perspective, the first developments in hierarchical clustering date back to the work of Sokal et al. in the field of numerical taxonomy, where systematic methods for evaluating similarity relationships between objects were proposed. Subsequently, this line of research was consolidated with reference works, such as Kaufman et al., who systematized the algorithms and validation metrics for cluster analysis. For its part, the heatmap technique has its roots in the graphic innovations introduced by Jacques Bertin in the field of visual semiotics and gained popularity in bioinformatics in the 2000s when combined with dendrograms generated by hierarchical clustering. A more recent and comprehensive review of this historical evolution can be found in the work of Wilkinson et al., who highlight how the integration of both methods gave rise to the so-called cluster heatmap. This tool is now a standard in multivariate analysis.

As a complementary technique to the exploratory analysis performed using PCA and the HJ-Biplot graph, a hierarchical clustering analysis was applied to the same standardized matrix to segment age groups based on their multivariate educational attendance profiles. This technique enables similar observations to be grouped based on a distance measure, resulting in a hierarchical structure of relationships that can be visually represented through a dendrogram.[28]

The clustering method selected was the *complete linkage* method, which considers the maximum distance between elements of two clusters when merging them, and Euclidean distance, appropriate for standardized numerical variables, was used as a measure of dissimilarity. This choice enables conservative segmentation, which tends to form compact and well-differentiated groups, making them suitable for analyzing educational profiles.

To implement this technique, the R programming language (version 4.4.2) was used again, which provides a series of specialized tools for cluster analysis. In particular, the dendextend package was used, an R library oriented towards the manipulation, customization, and visual analysis of dendrograms. This package allows not only the creation of clustering trees, but also their graphical modification, comparison of different clustering solutions, and adjustment of the visual styles of branches, labels, and nodes. In this study, dendextend played a key role in visually organizing the results of hierarchical clustering and highlighting the relationships between different age groups.

In addition to the dendrogram, a heatmap was constructed to simultaneously observe the intensity of the variables by age group and the similarity structure between observations. The pheatmap package, designed specifically for generating heatmaps in R, was used for this purpose. This package combines matrix data representation with dendrograms associated with both rows and columns, facilitating the integrated interpretation of high- and low-value patterns, as well as the structural proximity between groups.

pheatmap allows you to adjust aspects such as color scales, labels, additional annotations, margins, and cell size, providing flexibility to produce a clear and academically appropriate visualization. In this study, a heat map was used to reinforce the results of hierarchical clustering visually and to verify whether the identified groups had homogeneous profiles with respect to the analysis variables (attendance, non-attendance, and population by sex).

Finally, to evaluate the quality of the clustering scheme, two internal validation metrics were calculated using R's own functions: the cophenetic correlation coefficient, which reached a value of 0,91, indicating high fidelity between the dendrogram and the original distances, and the average silhouette index, with a value of 0,61, suggesting that the cluster structure obtained is reasonably well defined, with acceptable separation between groups.

The integration of these tools not only facilitated the visual exploration of educational behavior by age but also allowed for the empirical validation of the existence of age segments with differentiated profiles, thereby strengthening the robustness of the multivariate analysis and providing evidence for the targeted delivery of educational interventions.

Evaluation of the quality of the grouping

Two internal validation metrics were calculated:

- Cophenetic correlation coefficient: 0,91 (high fidelity of the dendrogram).
- Average silhouette index: 0,61 (acceptable clustering structure).

Tools used

All analysis was performed in the R programming environment (version 4.4.2), using the following packages:

- FactoMineR and factoextra (PCA and HJ-Biplot)
- ggplot2 (visualization)
- dendextend and pheatmap (clustering and heatmap)

This study was conducted in accordance with the ethical principles established for scientific research in the social sciences and humanities, respecting at all times the integrity of the data, the responsible use of information, and the rights of individuals, even though no data was collected directly.

As this analysis is based on secondary, aggregated, and anonymized data from the VIII Population Census and VII Housing Census 2022 of the National Institute of Statistics and Census (INEC) of Ecuador, no informed consent or direct intervention with human subjects was required. The information used is publicly and freely available, in accordance with the transparency and statistical dissemination policies established by INEC.

In accordance with current international and national regulations on research ethics (such as the Declaration of Helsinki, the Belmont Principles, and the Codes of Ethics of Latin American universities), this study:

- Does not violate the confidentiality or privacy of individuals or communities, as it does not use microdata or identify specific individuals.
- Does not involve physical, psychological, or social risks for the subjects, since the information is purely statistical and aggregated.
- Respects the right of open access to information for academic purposes, in line with the principle of democratization of knowledge.
- Guarantees the accuracy of the results and the correct citation of all sources used, in accordance with criteria of scientific integrity.

Furthermore, it is hereby declared that:

- There is no conflict of interest on the part of the authors.
- No external funding was received.
- The work was carried out under standards of methodological transparency, reproducibility, and peer review.

Consequently, the study complies with the ethical requirements for research using public secondary data, and its development does not contravene any applicable institutional, national, or international regulations.

RESULTS

Before presenting the multivariate findings, a brief demographic description of the population analyzed is provided below. The sample consists of 27 five-year age groups, from 0–4 years to 85 years and older, segmented by sex. This classification allows us to observe the evolution of educational attendance throughout the life cycle.

According to data from the VIII Population Census and VII Housing Census[1], the total population analyzed amounts to approximately 17,5 million people, comprising 8,9 million women and 8,6 million men. The largest groups correspond to ages 5–19, coinciding with the basic and secondary education cycles. The population declines progressively from age 40 onwards, with a more marked decline in the age groups above 65.

In terms of educational attendance, the data reflect a high concentration of attendance in the 5- to 19-year-old age groups, peaking in the 10-to 14-year-old age group, where more than 90 % of people attend an educational institution. From the age of 20 onwards, attendance declines progressively, with significant drops in the 30 to 49 age groups. In these groups, a considerable number of people are identified as not attending school, suggesting a cumulative phenomenon of school lag or dropout. The age groups over 60 have very low attendance, reflecting historical trends of limited access to formal education in previous generations.

Factor composition of the HJ-Biplot

Principal component analysis revealed that the first two components explain 99,96 % of the total variance, with the first (PC1) accounting for 79,5 % and the second (PC2) for 20,4 %. This high explanatory power justifies the use of a two-dimensional factorial plane to represent all the relevant information contained in the data.

The HJ-Biplot graph allowed the vectors of the original variables and the age groups to be represented simultaneously as observations. It was observed that the total male and total female variables align positively with the PC1 axis. In contrast, the Total_Attendance variable has a projection closer to the PC2 axis, indicating a lower contribution to the main structure that defines population differences in terms of non-attendance at school.

The age groups between 30 and 49 years, especially those between 30-34, 35–39, and 40–44 years, were located at the extreme ends of the PC1 axis, indicating a strong association with the Total_No_Attendance variable. This location in the factorial space suggests a close link between these groups and the variable "Total number of people who do not attend," suggesting a clear association with higher levels of educational exclusion in the adult population. Their position at the extreme ends of the central axis reinforces the interpretation that these age ranges account for a significant part of the structural educational gap.

This location in the factorial space indicates a close link between these groups and the variable "Total number of people who do not attend," suggesting a clear association with higher levels of educational exclusion in the adult population. Their position at the extreme ends of the central axis reinforces the interpretation that these age groups account for a significant portion of the structural educational gap.

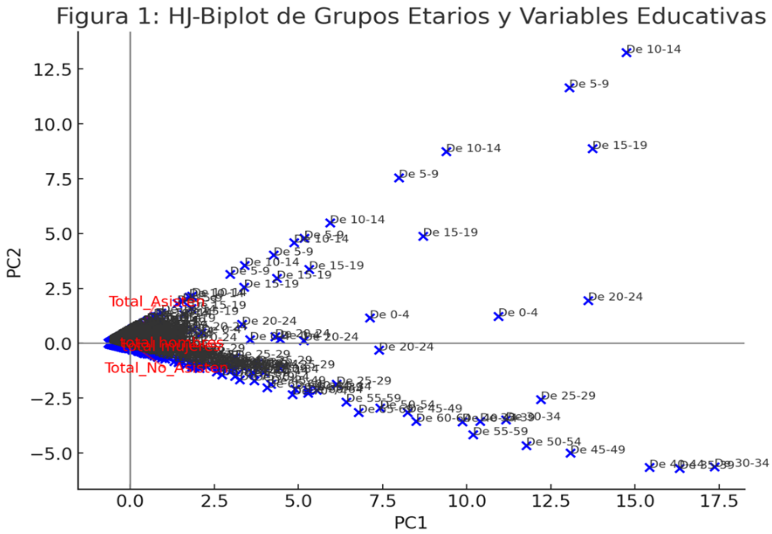

Figure 1. Hj-Biplot of Age Groups and Educational Variables

HJ-Biplot representation of age groups and educational variables in Ecuador, 2022 Census

The HJ-Biplot graph presented in figure 1 was generated automatically using the R programming language (version 4.4.2), specifically through the FactoMineR, factoextra, and ggplot2 packages, which allow the development of visual representations in vector and editable format. The database used to create this figure corresponds to the results of the VIII Population Census and VII Housing Census 2022, published by the National Institute of Statistics and Censuses (INEC) of Ecuador. The information was extracted from the national table on attendance at educational establishments, disaggregated by five-year age group and sex, available on the official INEC website.

Pareto classification and strategic visualization

Based on the Pareto analysis applied to the variable "Total number of people not attending," it was determined that only nine age groups (out of the 27 analyzed) account for 80 % of the national total of non-attendance in education. These correspond to the following age ranges: 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, and 60-64 years. These were classified as Group A, with the 30-49 age group standing out in particular, as it has the highest absolute prevalence of educational exclusion.

The groups between 65 and 84 years of age were categorized as Group B, representing an additional 15 % of non-attendance. Finally, the remaining age groups, mainly those of very young ages (0-4 and 5-9 years) and older ages (85 years and over), comprised Group C, accounting for only 5 % of the cumulative total.

This classification was incorporated into the HJ-Biplot graph using different colors (red for Group A, orange for Group B, and green for Group C), allowing for a clear visualization of the priority segments for intervention. The visual layout showed that Group A tends to concentrate in the right quadrant of the factorial plane, reinforcing its alignment with the variable with the most significant weight in the analysis: non-attendance at school.

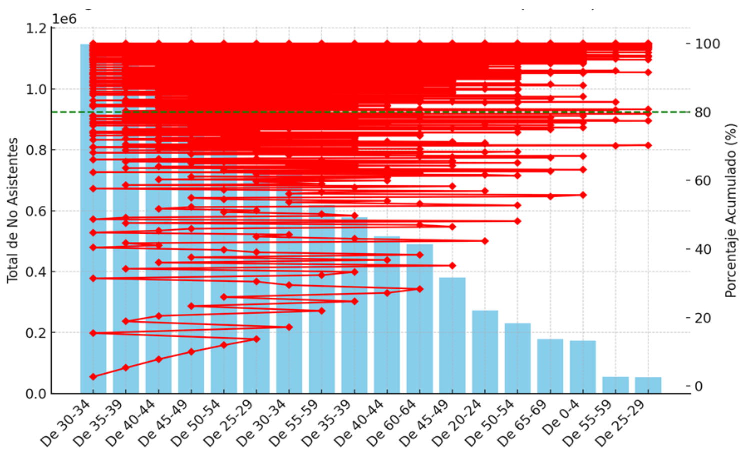

Figure 2. Pareto- graph - Educational Exclusion by Age Group

Pareto chart applied to the variable "Total number of people not attending school."

The graph enables age groups to be ranked according to their cumulative contribution to educational non-attendance, highlighting the segments considered a priority according to the Pareto principle (80/20 rule).

Figure 2 was generated automatically using the R programming language (version 4.4.2) and functions from the ggplot2 package, which enables the creation of vector- and editable-format graphs suitable for academic publications. This figure represents a Pareto chart applied to the variable "Total number of people who do not attend," analyzed from census data. The data source corresponds to the results of the VIII Population Census and VII Housing Census 2022, published by the National Institute of Statistics and Census (INEC) of Ecuador, specifically from the national table on attendance at educational establishments by five-year age group and sex.

Hierarchical clustering and complementary analysis

To validate and enrich the above findings, a hierarchical clustering analysis was applied to the standardized matrix. The resulting dendrogram showed three well-defined clusters:

- Cluster 1, comprising the 30-49 age group, was characterized by a homogeneous internal structure, with high levels of both non-attendance at educational establishments and population density.
- Cluster 2, comprising middle-aged and older groups (50 to 64 years), exhibits similar characteristics, albeit with less intensity.
- Cluster 3, comprising younger (5-19 years) and older (70+) age groups, exhibited a low association with non-attendance and differentiated levels of active attendance, reflecting its peripheral location in the HJ-Biplot.

These results were represented by a heatmap, which visualized not only the relative magnitudes of each variable by age group, but also their membership in each cluster. The convergence between the Pareto classification and the hierarchical clustering segmentation reinforces the validity of the proposed methodological approach.

The HJ-Biplot indicates that the first two components account for 99,96 % of the total variability. Variable vectors such as 'Total_No_Attendance' and 'total men/women' dominate the PC1 axis, while 'Total_Attendance' dominates the PC2 axis. Age groups classified as A (red) account for 80 % of non-attendance and should be considered a priority in educational strategies.

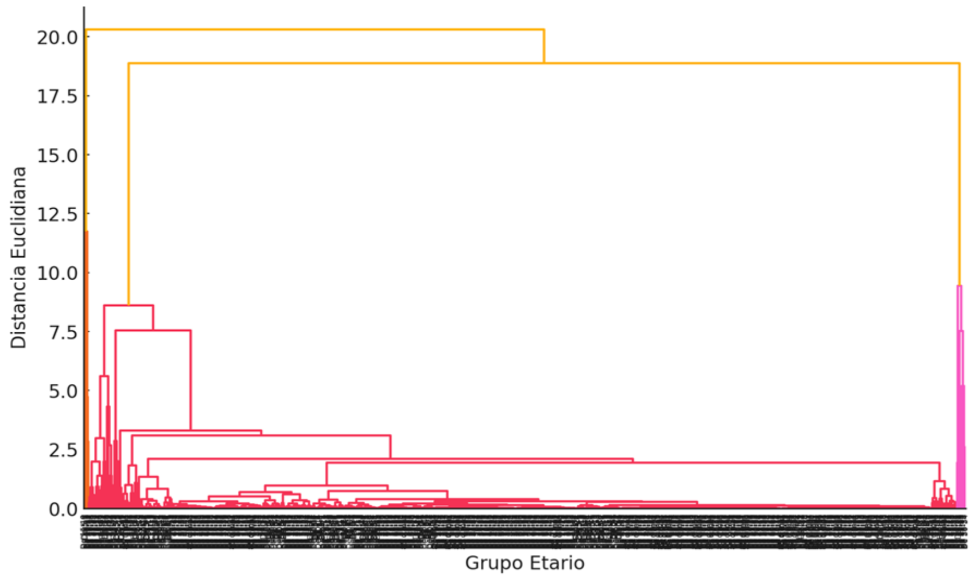

Figure 3. Dendrogram of Age Group Clustering

Hierarchical clustering dendrogram of age groups according to educational attendance patterns. The hierarchical structure shows the formation of three main clusters, highlighting similarities and differences between age profiles.

Figure 3 was generated automatically using the R programming language (version 4.4.2) and functions from the dendextend package, which specializes in the construction and customization of dendrograms in vector and editable formats. This figure corresponds to a hierarchical clustering analysis applied to the age groups in the study, based on their educational attendance profiles. The source of the data used is the VIII Population Census and VII Housing Census 2022, prepared by the National Institute of Statistics and Census (INEC) of Ecuador, based on the national table on attendance at educational establishments by five-year age group and sex.

Figure 4. Heatmap of Educational Variables by Age Group

Hierarchical clustering dendrogram of age groups according to educational attendance patterns. The hierarchical structure shows the formation of three main clusters, highlighting similarities and differences between age profiles.

Figure 3 was generated automatically using the R programming language (version 4.4.2) and functions from the dendextend package, which specializes in the construction and customization of dendrograms in vector and editable formats. This figure corresponds to a hierarchical clustering analysis applied to the age groups in the study, based on their educational attendance profiles. The source of the data used is the VIII Population Census and VII Housing Census 2022, prepared by the National Institute of Statistics and Census (INEC) of Ecuador, based on the national table on attendance at educational establishments by five-year age group and sex.

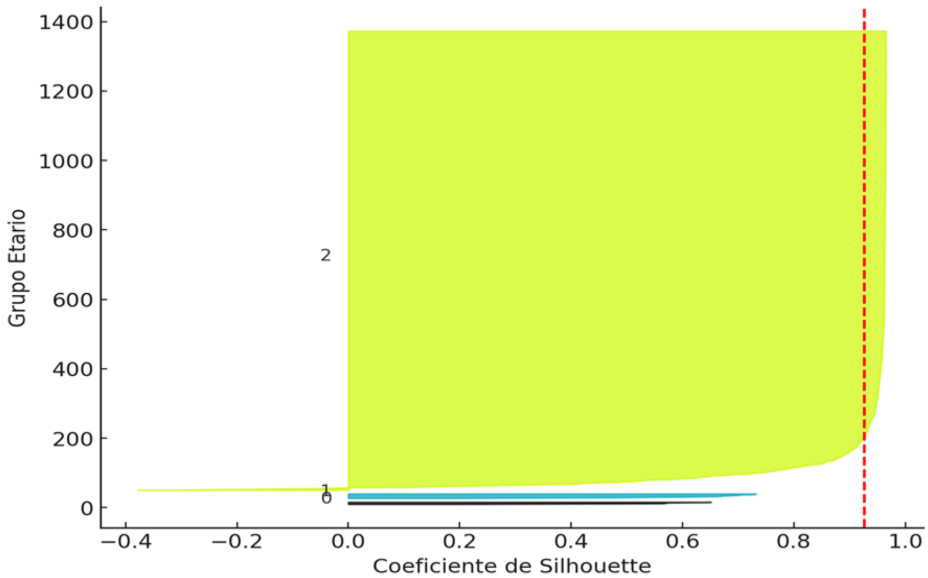

Figure 5. Silhouette diagram for 3 clusters

Silhouette graph for validation of hierarchical clustering of age groups.

The average silhouette coefficients confirm a reasonable and well-defined structure for the three clusters, validating the segmentation performed.

Figure 5 was generated automatically using the R programming language (version 4.4.2), utilizing functions from the cluster package, along with factoextra, both of which are designed for cluster analysis and evaluation.

The silhouette diagram enables the assessment of the internal quality of the clustering model, graphically representing the silhouette coefficient for each observation, which indicates the degree of membership in the assigned group. This figure was generated in vector and editable format, in accordance with scientific publication standards. The data used comes from the VIII Population Census and VII Housing Census 2022, published by the National Institute of Statistics and Census (INEC) of Ecuador, specifically based on the national table on attendance at educational establishments by five-year age group and sex.

DISCUSSION

The results of the multivariate analysis provide a solid basis for interpreting educational exclusion in Ecuador, from both a methodological and substantive perspective. The application of techniques such as HJ-Biplot, Pareto analysis, and hierarchical clustering enabled the representation of structural relationships between educational variables, segmenting age groups by common behaviors, and prioritizing interventions based on objective criteria.

HJ-Biplot identified a clear association between the 30-49 age group and the variable related to non-attendance in education. This finding highlights that educational exclusion is not restricted to childhood or adolescence but also affects adult cohorts with interrupted school trajectories. This situation has been documented in recent research, which links socioeconomic inequality with educational backwardness, particularly in contexts of persistent vulnerability.[8,9,10]

In line with this, the Pareto analysis identified nine age groups that account for the highest proportion of people outside the education system. This type of classification facilitates decision-making by defining priority segments. It has been utilized in current research to understand exclusion based on variables such as gender, household schooling, or environmental conditions.[11,12]

The incorporation of hierarchical cluster analysis and its representation in a heat map served as a technique for validating the findings. The coherence between the identified clusters and the Pareto classification supports the internal consistency of the analysis, reinforcing the methodological value of combining multivariate techniques. This type of approach has been applied in education systems where social inequalities tend to be reproduced through structural segmentation mechanisms.[13,14]

From a methodological perspective, the combined use of principal component analysis, HJ-Biplot visualization, Pareto classification, and hierarchical clustering provides a robust framework for diagnosing educational inequalities. This integration facilitates the construction of differentiated profiles of exclusion and can serve as a basis for public policies aimed at educational reintegration, especially among the adult population. In addition, recent research highlights the use of advanced techniques such as machine learning to predict school dropout and support educational planning in various contexts.[15]

In summary, the findings support the use of multivariate approaches as a valuable analytical tool in the study of complex educational phenomena and provide an empirical typology that can inform differentiated strategies based on the exclusion profiles identified in the Ecuadorian context.

On the other hand, the adoption of emerging technologies for inclusive purposes, explored in sectors such as sustainable finance,[16] has demonstrated its potential to reduce structural barriers. This approach can be transferred to the educational sphere through personalized digital platforms that facilitate the re-entry of adults into the educational system, especially in contexts where exclusion is due to economic or geographical limitations.

Although the analysis was based on aggregate national census data, the proposed methodology has a wide range of applications. It can be adapted to other social areas, such as health, employment, or poverty, at more specific territorial scales (province, canton, or urban and rural areas), or even applied to future data series that allow for longitudinal analyses of educational backwardness.

The use of multivariate surveys to study social perceptions of sustainability, as proposed in [17] is also helpful in analyzing educational perceptions in highly vulnerable contexts.[18] show how the urban-rural geographical environment has a significant impact on educational gaps, through the use of machine learning algorithms applied to academic performance in Ecuador.

For their part [19] applied Disjoint HJ-Biplot clustering techniques to identify patterns of educational exclusion in Latin American contexts and concluded that this tool offers greater visual accuracy for classifying social groups compared to traditional methods.

In turn, recent research on structural discrimination in vulnerable populations has highlighted the usefulness of multivariate analysis in identifying non-obvious patterns of social exclusion, as noted by [20] in their study on accumulated inequalities.

Methodological innovation of the study

This study presents an innovative approach to analyzing educational phenomena by systematically integrating HJ-Biplot, Pareto analysis, and hierarchical clustering with official census data. Although each of these techniques has been used individually in academic or social contexts, their strategic combination in a single analytical framework to address educational exclusion in Latin America has not been widely reported in recent literature. This integration enables not only the visualization of complex patterns but also the simultaneous prioritization of interventions and segmentation of population profiles, providing a replicable methodological tool for studies of educational inequality and other multivariate phenomena. In this way, this work makes an original methodological contribution that can enrich the statistical analysis applied to public policies in the region.

Limitations of the study

Despite the robustness of the methodological approach used, this study has some limitations that should be considered. First, the analyses were based on census data aggregated by five-year age groups, which makes it impossible to capture intra-group variations, individual heterogeneities, or additional contextual factors that could influence educational exclusion, such as socioeconomic status, specific geographic location, or ethnicity. Likewise, the cross-sectional nature of the data prevents the observation of evolutionary dynamics over time, limiting the analysis of educational trajectories. Finally, although the combined use of HJ-Biplot, Pareto analysis, and hierarchical clustering offers a robust approach for identifying patterns, future research could broaden the approach by incorporating additional variables and exploring longitudinal or multilevel methodologies that allow for greater capture of future research could broaden the approach by incorporating additional variables and exploring longitudinal or multilevel methods, also considering phenomena of spatial segregation that affect educational performance, such as those analyzed by (21).

Practical implications

The results of this study offer a strategic tool for decision-making in public education policy in Ecuador. The identification of patterns of educational exclusion using multivariate techniques enables the prioritization of vulnerable territories and population groups, while also considering the socioeconomic factors that affect basic digital skills in different European demographic groups.(22) In addition, the visualization methods used facilitate the communication of complex findings to non-technical audiences, supporting the formulation of targeted intervention programs and the design of sustainable educational equity strategies. Similarly (23) identifies specific school factors, such as classroom environment and teacher support, that enhance the performance of socioeconomically disadvantaged students, thus offering concrete clues for the design of inclusive policies.

CONCLUSIONS

The study concludes that educational exclusion in Ecuador follows structural patterns linked to specific age groups, which justifies the need for differentiated and targeted hy strategies. The integration of multivariate techniques, such as HJ-Biplot, Pareto analysis, and hierarchical clustering, facilitates a comprehensive understanding of the phenomenon by offering visual and interpretive representations of the dynamics of educational attendance and non-attendance.

The methodological proposal presented is configured as an adaptable tool with the potential to be applied to different social dimensions and territorial scales. This flexibility enables the planning of more equitable, evidence-based public policies that are tailored to local realities. Along these lines, proposes a model that allows the mapping of human values' diversity using artificial intelligence, which opens up new possibilities for contextualizing educational objectives within ethical frameworks that are sensitive to cultural and social differences.

Likewise, emphasizes the importance of establishing clear indicators and encouraging the active participation of various social actors as necessary conditions for designing sustainable education strategies. This approach complements the proposal of this study by highlighting the need to consider both structural and participatory elements when defining educational inclusion policies.

Future lines of research include the design of longitudinal studies to analyze the evolution of educational exclusion over time. It would also be relevant to apply multilevel analysis models that capture the dynamics between individuals, academic institutions, and regions. Another relevant avenue of analysis would be to compare educational exclusion in urban and rural contexts using multivariate time series approaches, which would allow for the identification of differentiated trajectories of educational lag and opportunity.

REFERENCES

1. Instituto Nacional de Estadística y Censos (INEC). VIII Censo de Población y VII de Vivienda. Quito: INEC; 2022.

2. UNESCO. Reimagining our futures together: A new social contract for education. Informe de la Comisión Internacional sobre los Futuros de la Educación. París: UNESCO; 2021.

3. UNESCO. Global Education Monitoring Report 2023: Technology in education – A tool on whose terms? París: UNESCO; 2023.

4. Agasisti T, Antequera C, Delprato M. Technology and equity in Latin American education: Lessons from PISA. OECD Education Working Papers. 2023.

5. Bonilla-Mejía L, Londoño-Ortega B. Brechas educativas y aislamiento geográfico: Evidencia desde zonas rurales en América Latina. Rev Latinoam Educ Comp. 2024;8(1):45-63.

6. Radtke J. Understanding the complexity of governing energy transitions: Introducing an integrated approach of policy and transition perspectives. Environ Policy Gov. 2025. https://doi.org/10.1002/eet.2158

7. Galego D, Lotta G, Baker K. Exploring the nexus of common good and the Neo-Weberian State: A comparative analysis of Brazil and Mexico. Public Adm Dev. 2025. https://doi.org/10.1002/pad.2105

8. Batool SM, Liu Z. Exploring the relationships between socio-economic indicators and student enrollment in higher education institutions of Pakistan. PLoS One. 2021;16(12):e0261577. https://doi.org/10.1371/journal.pone.0261577

9. Kaffenberger M. Modelling the long-run learning impact of the COVID-19 learning shock: Actions to (more than) mitigate loss. Int J Educ Dev. 2021;81:102326. https://doi.org/10.1016/j.ijedudev.2020.102326

10. Shores K, Kim HE, Still M. Categorical inequality in Black and White: Linking disproportionality across multiple educational outcomes. Am Educ Res J. 2020;57(2):556-97. https://doi.org/10.3102/0002831219900128

11. Akram S, Hassan MU. A multilevel analysis of out-of-school children by gender and educational level. Child Soc. 2025;39(2):123-39. https://doi.org/10.1111/chso.12933

12. Entrich SR. Worldwide shadow education and social inequality: Explaining differences in the socioeconomic gap in access to shadow education across 63 societies. Int J Comp Sociol. 2020;61(4):271-93. https://doi.org/10.1177/0020715220987861

13. Schmidt WH, Burroughs NA, Cogan LS. The role of schooling in perpetuating educational inequality: An international perspective. J Curric Stud. 2020;52(4):519-39.

14. Domina T, Penner A, Penner E. Categorical inequality: Schools as sorting machines. Annu Rev Sociol. 2020;46:311-29.

15. Sansone D. Predicting school dropout with machine learning: Evidence from Chile. Econ Educ Rev. 2021; 82:102125.

16. Barua S, Golder U, Chowdhury RS. Implications of NFT as a sustainable fintech innovation for sustainable development and entrepreneurship. Sustain Technol Entrep. 2025;4:100092. https://doi.org/10.1016/j.stae.2025.100099

17. Helen Vaikma. Unveiling the mindset: measuring consumer perception towards the dimensions of sustainability. Sustainable Futures. 2025;9(2):100616. https://doi.org/10.1016/j.sftr.2025.100616.

18. Santos-García F, Valdivieso KD, Rienow A. Urban–rural gradients predict educational gaps: Evidence from a machine learning approach. ISPRS Int J Geo-Inf. 2021;10(12):830. https://doi.org/10.3390/ijgi10120830

19. Mendoza-Pérez J, Álvarez-Castro C, Ríos-Bustamante A. Multivariate clustering and educational exclusion: A Latin American approach using HJ-Biplot. Rev Latinoam Estud Educ. 2022;52(1):45-68.

20. Reisner SL, Johnson N, Chen JT, Marini M, LeBlanc ME, Mayer KH. Analyzing multiple types of discrimination using implicit and explicit measures, comparing target vs. dominant groups, in a study of smoking/vaping among marginalized populations. Int J Equity Health. 2025;24(1):456-72. https://doi.org/10.1186/s12939-025-02456-9

21. Otero G, Carranza R, Contreras D. Spatial divisions of poverty and wealth: Does segregation affect educational achievement? Socioecon Rev. 2023;21(1):617-46. https://doi.org/10.1093/ser/mwab022

22. Fodranova I, Labudova V. The impact of socio-economic factors on digital skills in the population of the EU 27 countries. Virtual Econ. 2024;7(1):29-45. https://doi.org/10.34021/ve.2024.07.03(5)

23. Agasisti T, Avvisati F, Borgonovi F. What school factors are associated with the success of socio-economically disadvantaged students? An empirical investigation using PISA data. Soc Indic Res. 2021;155(1):271-97. https://doi.org/10.1007/s11205-021-02668-w

24. Zhang J, López D, Jiménez R. Using multivariate analysis for educational inequality diagnostics in Latin America. Educ Policy Anal Arch. 2022;30(40):1-24. https://doi.org/10.48550/arXiv.2504.09861

25. Basheer N, Ahmed V, Bahroun Z, Anane C. Sustainability assessment in higher education institutions: exploring indicators, stakeholder perceptions, and implementation challenges. Discov Sustain. 2025;6(1):116. https://doi.org/10.1007/s43621-025-01116-w

26. Galindo Villardón MP. Una alternativa de representación simultánea: HJ Biplot. Qüestiió. 1986;10(1):13-23. http://eudml.org/doc/40045

27. Pareto V. Cours d'économie politique. Lausanne: F. Rouge; 1896.

28. Wilkinson L, Friendly M. The history of the cluster heat map. Am Stat. 2009;63(2):179-84. doi:10.1198/tas.2009.0033

FUNDING
The authors did not receive financial support for the development of the research, authorship, or analysis of the results. However, if the publisher accepts it, Milagro State University (UNEMI) will provide funds to cover the costs of publication.

CONFLICT OF INTEREST
The authors declare that there are no conflicts of interest related to the research, authorship, and/or publication of this article.

AUTHOR CONTRIBUTION
Conceptualization: Ángel Antonio Morán-Herrera, Melissa Mercedes Idrovo-Hurel, Dennis Alfredo Peralta-Gamboa.
Data curation: Angel Antonio Moran-Herrera.
Formal analysis: Dennis Alfredo Peralta-Gamboa.
Research: Angel Antonio Moran-Herrera, Melissa Mercedes Idrovo-Hurel.
Methodology: Ángel Antonio Morán-Herrera, Dennis Alfredo Peralta-Gamboa.
Project management: Dennis Alfredo Peralta-Gamboa.
Resources: Melissa Mercedes Idrovo-Hurel.
Software: Dennis Alfredo Peralta-Gamboa.
Supervision: Melissa Mercedes Idrovo-Hurel.
Validation: Ángel Antonio Morán-Herrera, Dennis Alfredo Peralta-Gamboa.
Visualization: Ángel Antonio Morán-Herrera.
Writing - original draft: Ángel Antonio Morán-Herrera.
Writing - review and editing: Melissa Mercedes Idrovo-Hurel, Dennis Alfredo Peralta-Gamboa.