engineering proceedings

MDPI

Proceeding Paper

A Hybrid Machine Learning and Monte Carlo Simulation Model for Educational Demand Forecasting in Ecuador †

Fernando Javier Pullutasig Acosta

Facultad de Ciencias Económicas, Administrativas y Negocios, Universidad Tecnológica Indoamérica, Ambato 180101, Ecuador; fernandopullutasig@uti.edu.ec

† Presented at the Sixth International Conference on Information Systems and Software Technologies (ICI2ST 2026), Quito, Ecuador, 23–25 September 2026.

Abstract

Educational planning in Ecuador relies on deterministic projections that fail to quantify demographic uncertainty. This study aimed to develop and validate a hybrid model integrating Machine Learning (XGBoost) and Monte Carlo simulation for probabilistic forecasting of educational demand at two levels of analysis: institution and canton × educational level. Five official sources were integrated—11.2 million birth records and 2.3 million death records (1990–2024), 331,067 enrollment records (2009–2024), international migration statistics (ESI), and the 2022 Census—into a panel with demographic lags of 3 to 17 years; XGBoost was benchmarked against Random Forest and a persistence baseline using strict temporal validation over 1-, 3-, and 5-year horizons, and prediction intervals were generated through 10,000 Monte Carlo simulations. Lagged births were the dominant predictor (SHapley Additive exPlanations (SHAP) importance 5864; correlations 0.95–0.99); at the cantonal level and five-year horizon, XGBoost reduced the root mean squared error (RMSE) by 24% relative to persistence; 80% prediction intervals achieved 79.1% empirical coverage; projected demand falls 12% by 2029, with 196 of 221 cantons facing excess supply. The hybrid approach yields calibrated projections that support territorial planning of educational supply.

Keywords: educational demand; forecasting; machine learning; Monte Carlo simulation; XGBoost; SHAP; demographic transition; Ecuador

1. Introduction

Forecasting educational demand is a central challenge for public policy planning in countries undergoing demographic transition. The recent literature has shown that stochastic forecasting methods, which quantify uncertainty through probability distributions, consistently outperform traditional deterministic projections in terms of usefulness for decision making [1]. In the demographic field, since the classical mortality model of Lee and Carter [2], functional data models with time-series coefficients have made it possible to generate coherent probabilistic forecasts of mortality, fertility, and migration, which are then combined through Monte Carlo simulation to obtain population projections by age and sex. The case for probabilistic rather than deterministic population forecasts has long been established [3], and successive reviews of demographic forecasting document the consolidation of stochastic methods as the methodological standard [4]. This paradigm now underpins global practice through Bayesian probabilistic projections for all countries [5], later extended to international migration [6]. Most of this literature, however, operates at

Check for updates

Academic Editor: Erick Herrera

Published: 18 September 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

the national scale; forecasting for small areas—such as cantons—poses specific challenges of volatility, data scarcity, and structural change that remain an active research frontier [7].

In parallel, advances in Machine Learning (ML) have established gradient boosting models, particularly XGBoost [8], as the state of the art for structured tabular data. These algorithms consistently outperform classical statistical methods in prediction tasks involving heterogeneous variables, missing data, and non-linear relationships [9]. Recent benchmark studies confirm that gradient-boosted tree ensembles remain superior to deep neural networks on tabular datasets of this kind [10]. The interpretability of these models has improved substantially with the development of SHAP (SHapley Additive exPlanations) [11], which quantifies the marginal contribution of each variable to individual predictions.

However, the intersection between Machine Learning and demographic forecasting applied to educational planning remains scarcely explored. Although there are studies that use cohort projections to estimate future enrollment [12,13], and others that apply ML to educational problems such as student dropout [14], enrollment decisions [15], or the relationship between educational–demographic indicators and labor outcomes [16], few studies combine both paradigms in a unified framework that simultaneously produces point predictions and probabilistic intervals.

Recent research has addressed the relationship between demographic dynamics and educational demand from several perspectives. Shen, Raymer, and Hendy [12] developed a school enrollment forecasting model for the Australian Capital Territory that integrates demographic projections with education system data, demonstrating that lagged births and migration are the main determinants of future demand. Wang [13] applied a similar approach in China to predict preschool education demand under the three-child policy, finding that the sustained decline in birth rates will cause a demand peak before 2030 followed by a significant contraction.

In the context of developing countries, Aidarkhanova et al. [17] assessed the impact of demographic growth on educational infrastructure in Kazakhstan, identifying regions with a critical deficit of school places. Lv et al. [18] projected the demand for rural teachers in western China under population change, confirming that the decline in births generates excess capacity in primary education and deficits in preschool. Jing et al. [19], in a systematic review of AI models for enrollment forecasting, concluded that gradient boosting models are the most effective for educational tabular data, but pointed out the lack of studies integrating demographic variables with ML. Beyond education, the same demography-to-demand logic has been applied to project physician requirements under population growth [20], and lagged cross-correlations between socioeconomic development and the demand for higher education have been documented for Ukraine [21], reinforcing the generality of demand frameworks driven by lagged demographic variables.

At the regional level, Miller et al. [22] documented the Latin American demographic transition and its impact on education, noting that Ecuador is in a phase of sustained reduction of its school-age population. Glewwe and Kremer [23] identified demand-side factors, such as family income and educational cost, as key determinants of enrollment in developing countries, while Epple and Romano [24] modeled competition between public and private schools, a relevant factor for the Ecuadorian context where 20.2% of institutions are private.

Ecuador has experienced a significant demographic transformation that continues to impact its educational system. Data from the National Institute of Statistics and Censuses (INEC) [25] reveal a sustained decline in births: from 369,692 in 2008 to 217,986 in 2024, a 41% reduction in 16 years. This trend has translated into a contraction of national school enrollment, which peaked in the 2014–2015 school year at 4,777,874 students and has fallen to 4,167,318 in 2024–2025, a 12.8% decrease. Simultaneously, the number of active educational institutions has dropped from 28,696 in 2009–2010 to 16,215 in 2024–2025,

indicating a process of system consolidation. These trends are not uniform across territories or funding types, raising the need for predictive tools that operate at a disaggregated level.

Despite the abundant availability of administrative data in Ecuador, several aspects remain undetermined. First, the predictive capacity of lagged births by canton on school enrollment has not been empirically quantified—neither at the level of the individual institution nor at the level of aggregate territorial demand—nor has the nature of this relationship (linear or non-linear). Second, no studies compare the performance of ML models against traditional demographic projection methods for enrollment forecasting in the Ecuadorian context. Third, the quality of the prediction intervals that a hybrid ML–Monte Carlo approach can produce, measured by their empirical coverage, is unknown. Finally, the cantons and provinces at greatest risk of future mismatch between educational supply and demand have not been systematically identified.

The general hypothesis of this study is that a hybrid model integrating XGBoost and Monte Carlo simulation, fed with lagged demographic variables (births, deaths, and migration), produces—at the aggregate territorial level (canton × educational level)—enrollment predictions with a lower overall error than reference methods (an RMSE reduction of at least 20% relative to the persistence baseline at the five-year horizon, with a WAPE not exceeding 15%) and calibrated prediction intervals (empirical coverage of the 80% interval within the 75–85% range), while acknowledging that at the individual institution level enrollment is dominated by persistence. Four specific hypotheses are also formulated: (H1) lagged births are the variable with the highest SHAP importance, surpassing migration, mortality, institutional, and local competition variables; (H2) the XGBoost model reduces RMSE by at least 20% relative to the persistence baseline; (H3) the 80% prediction intervals generated by Monte Carlo show an empirical coverage between 75% and 85% at the five-year horizon; and (H4) there are significant differences in predictive accuracy between urban and rural institutions, with larger errors in rural areas with high migration variability.

The general objective of this research is to develop and validate a hybrid model integrating Machine Learning (XGBoost) and Monte Carlo simulation for the probabilistic forecasting of educational demand in Ecuador at two levels of analysis (institution and canton × educational level), using historical demographic and educational data, and to derive from it a medium-term (3–5 years) territorial projection for supply planning. The specific objectives are: (1) to build a unified two-level dataset integrating births, deaths, migration, and enrollment with temporal lags of 3 to 17 years; (2) to identify the most predictive variables through SHAP analysis; (3) to train and compare three predictive models using strict temporal validation over 1-, 3-, and 5-year horizons; (4) to generate 80% and 95% probabilistic prediction intervals and assess their empirical coverage; (5) to produce a territorial analysis projecting cantonal demand to 2027 and 2029; and (6) to evaluate urban–rural differences in enrollment stability and their association with migration.

The paper is organized as follows: Section 2 presents the methodological framework, including data, models, and analysis techniques. Section 3 reports the results. Section 4 discusses them in light of the literature and presents the limitations. Finally, Section 5 offers conclusions and future research directions.

2. Materials and Methods

2.1. Research Approach and Design

This study follows a quantitative approach with a non-experimental, predictive design, since it seeks to estimate future values of a variable (school enrollment) from empirical relationships identified in historical data, without manipulating independent variables. Its scope is correlational–predictive, aimed at determining the association between lagged demographic variables and educational demand, and at using that association to gen-

erate forecasts with explicit uncertainty quantification. The research is applied, as the results have direct utility for the planning of educational supply. Two complementary units of analysis are defined: the institution–school year and the aggregate demand per canton × educational level × year, a distinction that makes it possible to assess at which level the demographic signal operates.

The methodology is justified for three reasons. First, the available data are exclusively quantitative (counts of births, deaths, migratory movements, and enrollment). Second, the research problem requires numerical predictions with uncertainty intervals, which rules out qualitative approaches. Third, the data volume (more than 13 million records) demands computational analysis techniques that only advanced quantitative methods can provide.

2.2. Population and Data

The study population comprises all educational institutions in Ecuador registered in the Master File of Educational Institutions (AMIE) of the Ministry of Education (MINEDUC) [26]. No sampling is applied; the complete universe is used. The primary unit of analysis is the institution–school year, producing a panel dataset of 331,067 observations from 33,563 unique institutions across 16 school years (2009–2010 to 2024–2025). The number of active institutions ranges from 28,696 (2009–2010) to 16,215 (2024–2025). On this basis, a panel of aggregate demand per canton × educational level (Initial Education, Basic General Education—EGB, and Baccalaureate) is also built for the 221 cantons.

Five official secondary data sources are integrated:

- **Live Birth Statistical Registry (INEC):** 11,185,982 individual records for the 1990–2024 period (35 years), with 79 variables including sex, province and canton of birth, mother's age and education, weeks of gestation, and type of delivery. These data allow computing the lagged births per canton that feed the predictive model.
- **General Deaths Statistical Registry (INEC):** 2,262,196 individual records for 1990–2024 (35 years), with 125 variables including sex, age, cause of death (ICD-10), and province and canton of death. They are used to build cantonal mortality indicators as a proxy for territorial health conditions.
- **MINEDUC Administrative Records (AMIE):** 331,067 enrollment records per institution for 16 school years, with 74 unified columns including total enrollment and enrollment disaggregated by educational level and sex (22 levels × 2 sexes), number of teachers, funding type, area, modality, and geographic location. Key columns are 100% complete.
- **International Entries and Exits (ESI, INEC):** statistics for the 1997–2025 period, from which the country's annual international net migration balance is derived as a national context variable.
- **VIII Population and Housing Census 2022 (INEC) [27]:** cantonal migration indicators—internal in-migration rate (habitual residence five years before the census) and international emigration per canton and year of departure.

Geographic harmonization across sources was performed with INEC's Statistical Geographic Classifier (DPA), building a crosswalk of 221 cantons that reconciles the codes of the different sources; in particular, the change of the canton code from 2 to 4 digits that occurred in 2011 was documented and resolved.

2.3. Data Engineering

The construction of the analytical dataset constitutes the first phase of the study and comprises four processes. First, birth aggregation: from the individual records, annual counts per canton and sex are computed, together with maternal profile indicators (proportion of adolescent mothers, average age). Second, death aggregation: counts are built per

canton, year, and age group (infant 0–4, school-age 5–17, general). Third, the construction of temporal lags connecting past births with future educational demand, covering the full age range of each level: lags of 3–4 years for Initial Education, 5–14 years for Basic General Education, and 15–17 years for Baccalaureate, whereby the demographic cohort of each level is built as the sum of births corresponding to the theoretical ages of that level. Fourth, aggregation of the demand panel: institutional enrollment is summed per canton × level × year and linked to its birth cohort, cantonal mortality, and migration.

For each observation, the final dataset includes lagged demographic variables (birth cohort per canton according to educational level), socioeconomic variables (lagged maternal profile), cantonal mortality, institutional variables (funding type, area, modality), local competition variables (number of institutions in the same parish and canton), migration variables (internal in-migration rate and international emigration from the 2022 Census; national migration balance from ESI), and autoregressive temporal variables (previous-period enrollment, used only in the "anchored" specification). The target variable is the total number of enrolled students.

Two exclusions were adopted for validity reasons. The year-over-year percentage change in enrollment was excluded from modeling because it constitutes information leakage (it derives directly from the target variable). The number of teachers was excluded from the demand analysis because it is a contemporaneous proxy of institutional size that masks the demographic signal to be isolated.

2.4. Predictive Models

Three models of increasing complexity are implemented. The persistence baseline assumes that enrollment in the target period will equal that of the origin period; any proposed model must outperform this reference to justify its complexity. The Random Forest model [9] provides a robust intermediate comparator. The XGBoost model [8] is the main model, selected for its documented superior performance on tabular data and its native capacity to handle missing values.

Each ML model is trained in two specifications: "anchored" (includes origin-period enrollment as an autoregressive variable) and "unanchored" (only demographic, institutional, and migration variables), which allows isolating the contribution of the pure demographic signal. XGBoost hyperparameters are optimized through Bayesian search with Optuna [28], using RMSE minimization on the validation set as the objective function. The search employs the Tree-structured Parzen Estimator (TPE) sampler with a fixed seed (42) and 30 trials per study; an independent study is run for each combination of forecasting horizon (1, 3, and 5 years) and specification (anchored/unanchored), that is, six studies and 180 trials in total for the cantonal demand model. Each trial trains up to 600 boosting rounds with early stopping after 40 rounds without improvement on the validation set, and the best configuration is then refit with up to 800 rounds under the same early-stopping rule; the search space is detailed in Table 1. The institution-level benchmark follows the same protocol with 20 trials per study and slightly wider bounds (maximum depth up to 10, minimum child weight up to 20). All experiments were executed on a consumer-grade laptop (AMD Ryzen 9 5900HX CPU, 8 cores/16 threads, 32 GB RAM, Windows 11) running Python 3.14 with XGBoost 3.2.0, scikit-learn 1.9.0, and Optuna 4.9.0; the complete cantonal modeling stage (hyperparameter search, final training, and SHAP analysis for the six studies) runs in approximately 4 min, and the Monte Carlo simulation (10,000 iterations per observation for the two evaluation horizons) in approximately 2.5 min; the end-to-end reproducibility of the pipeline was verified by re-executing it under the fixed seed, which regenerates models and coverage results identically. Model interpretability is analyzed with SHAP [11], which provides global variable importance and local per-prediction explanations.

Table 1. XGBoost hyperparameter search space explored with Optuna (Tree-structured Parzen Estimator (TPE) sampler, seed 42; 30 trials per study for the cantonal demand model). Remaining parameters keep their XGBoost 3.2 defaults, with objective = reg:squarederror.

Hyperparameter	Distribution	Range
max_depth	integer (uniform)	3–9
learning_rate	float (log-uniform)	0.02–0.3
subsample	float (uniform)	0.6–1.0
colsample_bytree	float (uniform)	0.6–1.0
min_child_weight	integer (uniform)	1–15
reg_lambda	float (uniform)	0.5–5.0

2.5. Temporal Validation

Strict temporal validation (time-based split) is used to respect the temporal structure of the data and avoid leakage of future information. The scheme divides the data into three sets: training (school years up to 2019–2020), validation (2020–2021 to 2021–2022, for hyperparameter optimization), and test (2022–2023 to 2024–2025, for final evaluation). No random cross-validation is used. Evaluation is performed at three forecasting horizons—1, 3, and 5 years—replicating the real use of the model in short- and medium-term planning.

Evaluation metrics include RMSE (root mean squared error, which penalizes large errors), MAE (mean absolute error, more robust to outliers), MAPE (mean absolute percentage error, computed on observations with enrollment $\geq$30 to avoid distortions in very small units), sMAPE, WAPE (weighted absolute percentage error, appropriate for aggregate planning), and R^2 (proportion of explained variance).

The emphasis given to each metric deserves justification, because percentage errors behave poorly when the denominator is small. In the institutional panel, the median enrollment is 56 students and 23% of institution–year observations have fewer than 20 students, so even small absolute deviations translate into very large percentage errors; this motivates computing MAPE only on units with enrollment $\geq$30 and interpreting it with caution. At the cantonal demand level, unit sizes span two orders of magnitude (10th percentile 176 versus 90th percentile 13,131 students), and the largest decile of canton–level units concentrates 68.8% of total enrollment. For planning applications we therefore prioritize WAPE, which weights each unit by its actual enrollment so that the metric reflects where students actually are, and RMSE, whose quadratic penalty is sensitive to large absolute misses in populous units—precisely those with the greatest budgetary and infrastructure consequences. MAPE and sMAPE are retained for comparability with the prior literature.

2.6. Monte Carlo Simulation

Monte Carlo simulation [29] is implemented as an extension of the XGBoost model to generate probabilistic prediction intervals. The procedure runs 10,000 iterations per observation, each consisting of three steps: (1) sampling the demographic input variables from their historical empirical distributions estimated per canton, (2) passing the perturbed feature vector through the trained XGBoost model to obtain an enrollment prediction, and (3) adding a residual noise term sampled via bootstrap [30] from the model's residuals on the validation set. This approach, inspired by the philosophy of Hyndman and Booth [1] of producing coherent probabilistic forecasts, guarantees that the simulated scenarios are consistent with the patterns learned from the data.

The analytical products of the simulation include prediction intervals at the 10–90 (80%) and 2.5–97.5 (95%) percentiles and the verification of their empirical coverage on the test set, that is, the proportion of observed values that actually fall within the nominal interval (H3).

A structural assumption of this procedure is the stationarity of input uncertainty: the empirical distributions of the demographic variables, estimated from the 1990–2024 record, are taken as representative of future variability—an assumption that is standard in stochastic demographic forecasting [1,2]. If future variability departs substantially from the historical range (an unprecedented health crisis, an abrupt change in migration policy, or an economic shock affecting fertility decisions), the intervals may under- or overestimate true uncertainty. Two mitigations follow. First, periodic re-estimation of both the model and the input distributions as each new annual cut of the registries becomes available; second, complementary stress-testing with scenarios beyond the historical range (for example, simulating birth declines steeper than any observed). The 35-year span of the demographic record—which covers several economic cycles, the 1999 crisis and subsequent dollarization, the COVID-19 pandemic, and the recent migration waves—provides a comparatively robust empirical basis, but it cannot anticipate genuinely novel disruptions.

2.7. Territorial Projection

Educational demand for 2027 and 2029 is projected using the cohort-component method, which exploits a structural property of the problem: the students who will enroll in the next 3 to 17 years have, for the most part, already been born. The demographic cohort of each target year is built from observed births per canton and multiplied by the most recent (2024) enrollment rate of each canton–level combination. For births not yet observed that affect the 2029 projection of Initial Education (2025–2026 cohorts), they are assumed to remain at the 2024 level, a conservative assumption. Each canton is classified according to the projected change in its demand as excess supply (decreasing demand), stable, or supply deficit (increasing demand), and the results are mapped at the cantonal level.

2.8. Computational Tools

Processing is carried out entirely in Python. Data engineering uses Polars (1.41.2) and Pandas (3.0.2) for the manipulation of the more than 13 million records. Modeling employs XGBoost (3.2.0) and Scikit-learn (1.9.0) for training and evaluation, Optuna (4.9.0) for Bayesian hyperparameter optimization, and SHAP (0.52.0) for interpretability. The Monte Carlo simulation is implemented with NumPy (2.4.4) and SciPy (1.17.1), and cartography with GeoPandas on GADM 4.1 administrative boundaries.

2.9. Ethical Considerations

All data used are publicly accessible and come from official sources of the Ecuadorian State (INEC and the Ministry of Education). No identifiable personal data are processed; birth and death records are aggregated at the cantonal level before being incorporated into the model. AMIE codes identify institutions, not individuals. The study does not require informed consent as it uses aggregated secondary data in the public domain.

3. Results

3.1. Educational Demand Is Governed by Demography (H1)

At the canton × educational level, the correlation between observed enrollment and the corresponding birth cohort is very high: 0.949 in Initial Education, 0.995 in EGB, and 0.988 in Baccalaureate. In the XGBoost demand model (5-year horizon, unanchored specification), the birth cohort dominates the SHAP importance ranking, exceeding the second demographic predictor by a factor greater than four (Table 2). Migration variables (internal in-migration from the 2022 Census) appear next, above mortality and provincial effects. **H1 is confirmed.**

https://doi.org/10.3390/engproc2026153021

Table 2. SHAP importance of the main variables in the cantonal demand model (XGBoost, 5-year horizon, unanchored specification).

#	Variable	SHAP Importance
1	Birth cohort of the level	5863.9
2	Educational level: EGB (structural dummy)	1654.2
3	Annual births of the canton	1276.1
4	Internal in-migrants, last 5 years (2022 Census)	643.3
5	Educational level: Baccalaureate (structural dummy)	355.5
6	Educational level: Initial (structural dummy)	337.8
7	Total deaths of the canton	249.5
8	Deaths, ages 0–4	114.5

Educational level dummies encode the structural scale of each level and do not constitute demographic predictors. SHAP = SHapley Additive exPlanations; EGB = Basic General Education.

3.2. Model Comparison: The Level of Analysis Matters (H2)

At the individual institution level, enrollment is highly persistent: the baseline assuming constant enrollment reaches $R^2 = 0.985$ at one year, and no ML model outperforms it at any horizon (Table 3). This result is consistent with theory: an institution's enrollment is the product of territorial demand times each school's market share [24], so the demographic signal—shared by all institutions in the canton—is diluted in the idiosyncratic variation of each establishment.

At the cantonal demand level, the picture changes (Table 4). At the five-year horizon, XGBoost based exclusively on demography reduces RMSE by 23.5% (2007 versus 2623 for persistence), meeting the 20% threshold set in H2, with a WAPE of 9.8% (below the 15% established in the general hypothesis). The improvement is concentrated in the largest cantons, which dominate the RMSE. In contrast, in the short term (1–3 years) and in median error metrics (MAE, WAPE), persistence remains a hard reference to beat. The high MAPE of unanchored XGBoost (31.8%) reflects large relative errors in small cantons, where a few dozen students represent high percentages; for aggregate planning decisions, WAPE is the relevant percentage metric. **H2 is partially confirmed:** the 20% threshold is met at the cantonal demand level and five-year horizon (−23.5%), but not at the individual institution level or at short horizons; the contribution of the demographic model is territorial and medium-term.

Table 3. Model comparison at the individual institution level: RMSE per horizon (test 2022–2024). RMSE = root mean squared error; RF = Random Forest; ML = Machine Learning. In all three horizons, the best ML model shows a higher RMSE than persistence.

Horizon	Persistence (Baseline)	Best ML Model	R^2 Persistence
1 year	58.4	59.7 (RF, anchored)	0.985
3 years	102.0	104.5 (RF, anchored)	0.954
5 years	122.9	149.4 (RF, anchored)	0.934

Table 4. Model comparison at the cantonal demand level (canton × level), 5-year horizon (test 2022–2024). "Unanchored" = only demographic, institutional, and migration variables; "anchored" = includes origin-period enrollment. MAE = mean absolute error; MAPE = mean absolute percentage error; WAPE = weighted absolute percentage error. MAPE computed on demand ≥ 30. Bold highlights the best RMSE and R^2, both achieved by the unanchored XGBoost model.

Model	RMSE	MAE	MAPE %	WAPE %	R^2
Persistence (baseline)	2623	540	11.85	8.49	0.9905
Random Forest (unanchored)	3274	728	17.16	11.46	0.9852
XGBoost (unanchored)	**2007**	622	31.79	9.79	**0.9944**
Random Forest (anchored)	3046	568	12.64	8.94	0.9872
XGBoost (anchored)	2238	686	77.93	10.78	0.9931

https://doi.org/10.3390/engproc2026153021

3.3. Uncertainty Calibration (H3)

The prediction intervals generated by Monte Carlo show adequate empirical coverage, especially at the five-year horizon (Table 5): the nominal 80% interval covers 79.1% of the test observations (within the 75–85% target range) and the 95% interval covers 93.6%. At three years, the 80% interval is slightly undercovered (73.8%). **H3 is confirmed** at the five-year horizon, which is precisely the planning horizon for which the demographic model adds value.

Figure 1 illustrates the intervals at the five-year horizon in the four cantons with the largest EGB demand: in all four cases the observed demand of the three test years falls within the 80% band, and the width of the bands makes explicit the forecast uncertainty that a deterministic projection would conceal.

Table 5. Empirical coverage of Monte Carlo prediction intervals (10,000 simulations per observation; n = 1989 test observations per horizon). PI = prediction interval. Bold highlights the five-year horizon, at which the nominal coverage target is met.

Horizon	80% PI Coverage (Target 75–85%)	95% PI Coverage	Median 80% PI Width
3 years	73.8%	90.1%	497
5 years	**79.1%**	**93.6%**	877

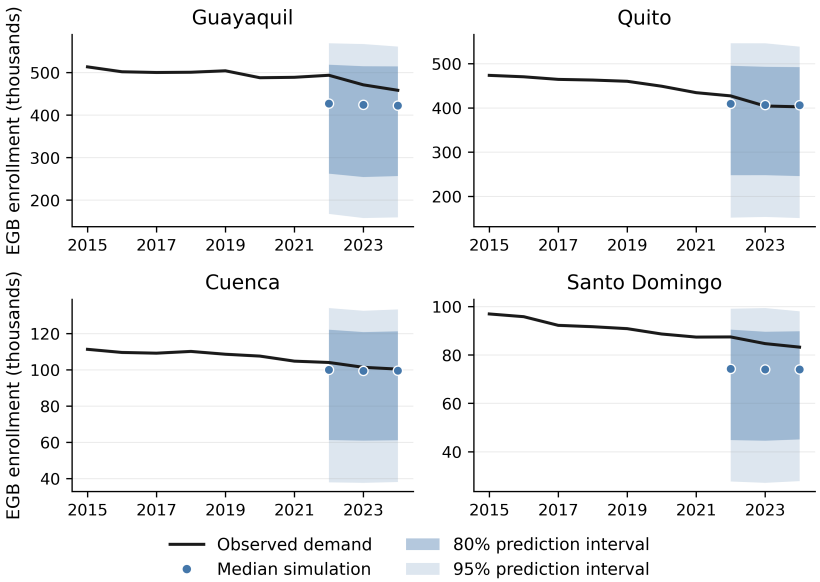


Figure 1. Monte Carlo prediction intervals (80% and 95%) versus observed demand at the five-year horizon, EGB level, for the four cantons with the largest demand (test years 2022–2024). Solid line: observed demand (2015–2024); dots: median simulation; dark band: 80% interval; light band: 95% interval. Source: own elaboration.

3.4. Urban–Rural Differences (H4)

Contrary to the usual assumption of greater rural instability, the year-over-year volatility of enrollment is higher in urban institutions (11.93%) than in rural ones (9.93%) (Table 6); the Mann–Whitney test provides no evidence in favor of the hypothesis of greater rural volatility ($p \approx 1.0$). Additionally, volatility is not associated with the intensity of cantonal internal migration (correlation $= -0.033$): high-migration cantons do not show larger errors than low-migration ones. **H4 is rejected:** enrollment instability is not a rural phenomenon associated with migration; it appears in urban contexts, plausibly due to the higher turnover derived from competition and school choice, while rural demand is more captive.

Table 6. Year-over-year enrollment volatility by area (test 2022–2024, institutions with enrollment ≥ 30; $n = 30{,}589$). Volatility measured as the persistence relative error $|E_t - E_{t-1}|/E_t$. Mann–Whitney U (H: rural > urban), $p \approx 1.0$.

Area	*n*	Mean Error %	Median %	90th Percentile %
Urban	19,243	11.93	7.32	26.92
Rural	11,346	9.93	6.85	21.81

3.5. Territorial Projection of Demand to 2027 and 2029

The cohort-component projection indicates a contraction of national educational demand from 4,108,221 students (2024) to 3,616,201 (2029), that is, **−12.0%**, with differentiated declines by level: −16.2% in Initial Education, −13.2% in EGB, and −6.3% in Baccalaureate (Table 7). The gradient by level reflects the advance of the reduced cohorts: the post-2015 birth decline hits Initial Education first and will reach Baccalaureate in the next decade.

At the territorial level, of the 221 cantons, 196 show projected excess supply (decreasing demand; 3.63 million enrollment in 2024), 16 remain stable, and only 9 show a deficit (increasing demand; 95 thousand enrollment), associated with expanding urban poles and migration-attraction areas: La Concordia (+44.2%), Camilo Ponce Enríquez (+22.0%), Montecristi (+10.1%), Daule, Durán, and Samborondón, among others. At the opposite extreme, the largest projected declines are concentrated in rural cantons of the central Sierra and the northern Amazon (Sucumbíos −39.7%, Guamote −28.5%, Colta −25.4%, Alausí −24.7%). The full territorial pattern is shown in Figure 2.

Table 7. National projection of educational demand by level, 2024–2029 (cohort-component method). Source: own elaboration based on INEC (births) and MINEDUC (enrollment). Bold indicates national totals.

Level	2024	2027	2029	Change 2024–2029
Initial	316,037	278,641	264,744	−16.2%
EGB	2,934,508	2,706,908	2,547,639	−13.2%
Baccalaureate	857,676	843,208	803,813	−6.3%
Total	**4,108,221**	**3,828,757**	**3,616,201**	**−12.0%**

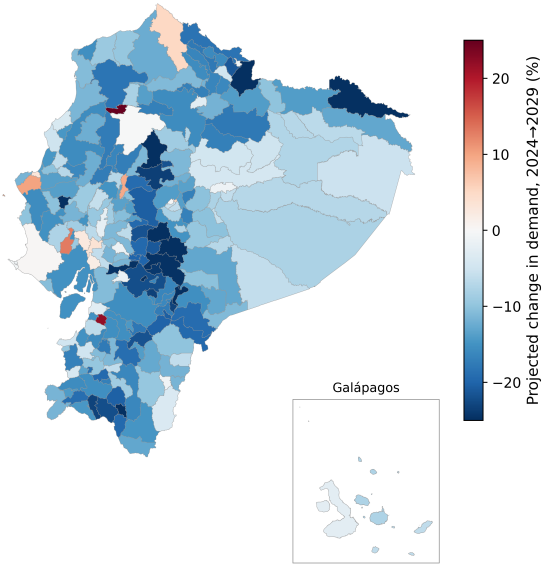


Figure 2. Projected change in educational demand by canton, 2024–2029 (Galápagos shown in the inset). Blue: excess supply (decreasing demand); red: supply deficit (increasing demand). Source: own elaboration with INEC and MINEDUC data; GADM 4.1 cartography.

4. Discussion

The central finding of this study is that demography is the structural driver of educational demand (H1), but its predictive power manifests at the aggregate territorial level and in the medium term, not at the individual institution level or in the short term (H2). This distinction—territorial population demand versus each establishment's market share—explains why persistence is practically unbeatable for predicting the enrollment of a specific school in the following year, and why the demographic model adds value in the aggregate five-year projection. Far from constituting a negative result, this nuanced reading is methodologically more solid than a uniformly superior ML performance, and it guides the correct use of each tool: persistence for short-term institutional management; the demographic model with Monte Carlo for territorial planning at 3–5 years.

The results engage directly with the recent literature. The primacy of lagged births coincides with the findings of Shen, Raymer, and Hendy [12] for the Australian Capital Territory, where births and migration determine future enrollment, and with Wang [13], who anticipates for China a contraction of preschool demand following the birth decline—the same pattern Ecuador already exhibits, with the particularity that here the contraction is documented with complete administrative records rather than fertility projections. The concentration of excess capacity in primary education projected by Lv et al. [18] for rural China finds its Ecuadorian analogue in the −13.2% decline in EGB and in the 196 cantons with excess supply. In turn, the conclusion of Jing et al. [19] on the superiority of gradient boosting for educational tabular data is confirmed here (XGBoost outperforms Random Forest in all demand configurations), but with the important precision that neither surpasses persistence at the institutional level: honest comparison against simple baselines, infrequent in that literature, changes the practical conclusions.

The contribution of the Monte Carlo component is the quantification of uncertainty with well-calibrated intervals at five years (H3), in line with the probabilistic forecasts that Hyndman and Booth [1] propose as the standard for planning. An 80% interval with 79.1% coverage allows the planner to size pessimistic and optimistic scenarios with quantified confidence, something traditional deterministic projections do not offer. The rejection of the classical urban–rural hypothesis (H4) also qualifies common assumptions of educational policy: enrollment instability is not rural nor explained by internal migration, but appears in competitive urban educational markets, consistent with models of competition and school choice [24].

From a public policy perspective, the projected 12% contraction of demand by 2029 and its territorial concentration have direct implications: resource reallocation across cantons, merger or conversion of establishments in the areas of greatest decline (central Sierra, northern Amazon), teacher workforce planning, and targeted investment in the few growth poles (the Guayas–Santa Elena corridor, migration-attraction cantons). The cantonal granularity of the model makes it possible to move from national diagnosis to localized decisions.

This study has limitations. First, the 2022 Census is a cross-sectional snapshot: internal migration is measured in a single window (2017–2022) and is assumed stable during the projection. Second, the 2029 projection for Initial Education assumes that births not yet observed (2025–2026) remain at the 2024 level, a conservative assumption which, if the declining trend continues, implies that the actual contraction would be larger than projected. Third, administrative boundaries changed during the study period (creation of provinces and cantons), which was resolved by harmonizing to a fixed delimitation, with a possible loss of historical detail. Fourth, MAPE penalizes small cantons, so accuracy conclusions rest on aggregate metrics (RMSE, WAPE) appropriate for planning but less informative about small individual units. Fifth, the model cannot anticipate demographic

shocks that fall outside the historical distribution of its inputs: events such as a new pandemic—COVID-19 already disturbs the 2020–2022 window used for validation and testing—could produce deviations that no interval calibrated on the past can fully capture. Sixth, changes in educational policy (modifications of compulsory schooling, subsidies, or rules governing private providers) can shift the mapping between demographic inputs and enrollment outcomes; the model represents institutional characteristics as features but does not model policy interventions as exogenous shocks. Seventh, economic fluctuations influence migration—as the recent Venezuelan immigration wave illustrates—and, although census-based and ESI migration variables are included, their future trajectories may differ from the historical patterns embedded in the training data. Finally, regarding external validity, the methodological framework (lagged demographic cohorts, gradient boosting benchmarked against persistence, Monte Carlo calibration) is transferable to any country with civil registration and enrollment records, but the estimated parameters, importance rankings, and cantonal projections are specific to Ecuador; validating the approach in other Latin American systems is a natural extension.

As future lines of research, we propose: explicitly modeling institutional market share (territorial demand → allocation among establishments), which would unite the two levels of analysis; contrasting the projections with INEC's official population projections as external validation; incorporating Monte Carlo uncertainty bands into the cantonal territorial projection; and extending the framework to teacher workforce and infrastructure planning.

5. Conclusions

First, it is feasible to build, entirely from public administrative records of Ecuador, a reproducible two-level analytical dataset (331,067 institution–year observations and a demand panel of 221 cantons × 3 educational levels) linking 35 years of births and deaths, migration, and 16 enrollment periods through demographic lags of 3 to 17 years (Objective 1).

Second, lagged births are the most important determinant of cantonal educational demand: they dominate the SHAP importance ranking (5864 versus 1276 for the second demographic predictor) and show correlations of 0.95–0.99 with enrollment by level, above migration, mortality, and institutional variables (Objective 2; H1 confirmed).

Third, the level of analysis conditions the value of Machine Learning: at the cantonal demand level and five-year horizon, XGBoost based solely on demography reduces RMSE by 24% relative to persistence (2007 vs. 2623; WAPE 9.8%), while at the individual institution level persistence dominates at all horizons ($\boldsymbol{R^2 = 0.985}$ at one year). The contribution of the demographic model is territorial and medium-term (Objective 3; H2 partially confirmed).

Fourth, Monte Carlo simulation (10,000 iterations) produces well-calibrated prediction intervals at the five-year horizon: empirical coverage of 79.1% for the nominal 80% interval (within the 75–85% target range) and 93.6% for the 95% interval (Objective 4; H3 confirmed).

Fifth, there is no evidence of greater enrollment instability in rural areas: year-over-year volatility is higher in urban institutions (11.93% vs. 9.93%) and is not associated with cantonal internal migration (correlation −0.033), shifting the attention of educational policy toward competitive urban markets (Objective 6; H4 rejected).

Sixth, the Ecuadorian educational system faces a structural contraction of demand: −12.0% by 2029 (from 4.11 to 3.62 million), with 196 of 221 cantons in excess supply and only 9 in deficit. Educational supply planning—reallocation, merger or conversion of establishments, and teacher workforce sizing—now has a probabilistic, reproducible tool with cantonal granularity to anticipate this adjustment (Objective 5).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study uses exclusively aggregated secondary data in the public domain, published by official sources of the Ecuadorian State (INEC and the Ministry of Education); no identifiable personal data were processed.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available from official public sources: live births and general deaths registries and the 2022 Population and Housing Census from the National Institute of Statistics and Censuses of Ecuador (https://www.ecuadorencifras.gob.ec (accessed on 3 July 2026)), and enrollment administrative records (AMIE) from the Ministry of Education of Ecuador (https://educacion.gob.ec/base-de-datos/ (accessed on 3 July 2026)). The processing pipeline is reproducible with a fixed global seed (42).

Acknowledgments: During the preparation of this manuscript, the author used Claude Opus 4.5 (Anthropic) for language translation and editing support. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Hyndman, R.J.; Booth, H. Stochastic population forecasts using functional data models for mortality, fertility and migration. *Int. J. Forecast.* **2008**, *24*, 323–342. [CrossRef]
2. Lee, R.D.; Carter, L.R. Modeling and Forecasting U.S. Mortality. *J. Am. Stat. Assoc.* **1992**, *87*, 659–671. [CrossRef]
3. Keilman, N.; Pham, D.Q.; Hetland, A. Why population forecasts should be probabilistic—Illustrated by the case of Norway. *Demogr. Res.* **2002**, *6*, 409–454. [CrossRef]
4. Booth, H. Demographic forecasting: 1980 to 2005 in review. *Int. J. Forecast.* **2006**, *22*, 547–581. [CrossRef]
5. Raftery, A.E.; Li, N.; Ševčíková, H.; Gerland, P.; Heilig, G.K. Bayesian probabilistic population projections for all countries. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 13915–13921. [CrossRef]
6. Azose, J.J.; Raftery, A.E. Bayesian Probabilistic Projection of International Migration. *Demography* **2015**, *52*, 1627–1650. [CrossRef]
7. Wilson, T.; Grossman, I.; Alexander, M.; Rees, P.; Temple, J. Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs. *Popul. Res. Policy Rev.* **2022**, *41*, 865–898. [CrossRef]
8. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
9. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
11. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30, Proceedings of NIPS 2017, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
12. Shen, T.; Raymer, J.; Hendy, C. Forecasting school enrollments in the Australian Capital Territory. *J. R. Stat. Soc. Ser. A* **2025**, *188*, 1107–1124. [CrossRef]
13. Wang, Y. Prediction of Preschool Education Demand under China's Three-Child Policy: Based on the Demographic Trends from 2024–2050. *Early Child. Educ. J.* **2025**, *53*, 1823–1834. [CrossRef]
14. Carballo-Mendívil, B.; Arellano-González, A.; Ríos-Vázquez, N.J.; Lizardi-Duarte, M.D.P. Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data. *Appl. Sci.* **2025**, *15*, 9202. [CrossRef]
15. Pham, T.-N.-D.; Nguyen, H.-D.-N.; Tran, T.-D. Analysis and prediction of enrollment decisions of university-admitted candidates. *Neural Comput. Appl.* **2025**, *37*, 26485–26499. [CrossRef]
16. Tleppayev, A.; Zeinolla, S. Forecasting Youth Unemployment Through Educational and Demographic Indicators: A Panel Time-Series Approach. *Forecasting* **2025**, *7*, 37. [CrossRef]
17. Aidarkhanova, G.; Zhumagulov, C.; Nyussupova, G.; Kholina, V. Assessing the Impact of Demographic Growth on the Educational Infrastructure for Sustainable Regional Development: Forecasting Demand for Preschool and Primary School Enrollment in Kazakhstan. *Sustainability* **2025**, *17*, 4212. [CrossRef]
18. Lv, Y.; Hao, B.; Zhang, X.; Wu, M. Demand structure prediction and adaptation path of rural teachers in western China under population change. *PLoS ONE* **2025**, *20*, e0331907. [CrossRef]
19. Jing, L.X.; Jasser, M.B.; Ajibade, S.-S.M.; Chua, H.N.; Wong, R.T.K. A Survey on Enrolment Analysis via AI-Based Forecasting Models in Academic Institutions. In *Proceedings of the 2025 IEEE 8th International Conference on Electrical, Control and Computer Engineering (InECCE), Kuantan, Malaysia, 27–28 August 2025*; IEEE: New York, NY, USA, 2025; pp. 654–659. [CrossRef]

20. Koichubekov, B.; Begaidarova, R.; Omarbekova, N.; Mukhanova, M.; Abdikadirova, K.; Kharin, A.; Omarkulov, B. Forecasting the impact of Kazakhstan population growth on healthcare doctors demand. *BMC Health Serv. Res.* **2025**, *25*, 1456. [CrossRef]
21. Kuryliak, V.; Lyzun, M.; Hayda, Y.; Lishchynskyy, I.; Ukhova, N. Cross-correlation analysis of dynamic interdependencies between socioeconomic development and the demand for higher education in Ukraine. *J. Eur. Econ.* **2025**, *24*, 467–485. [CrossRef]
22. Miller, T.; Saad, P.; Martínez, C. *La Nueva Era Demográfica en América Latina y el Caribe: La Hora de la Igualdad Según el Reloj Poblacional*; CELADE—División de Población, CEPAL: Santiago, Chile, 2014.
23. Glewwe, P.; Kremer, M. Schools, Teachers, and Education Outcomes in Developing Countries. In *Handbook of the Economics of Education*; Elsevier: Amsterdam, The Netherlands, 2006; Volume 2, pp. 945–1017. [CrossRef]
24. Epple, D.; Romano, R. Competition between Private and Public Schools, Vouchers, and Peer-Group Effects. *Am. Econ. Rev.* **1998**, *88*, 33–62.
25. Instituto Nacional de Estadística y Censos. Registro Estadístico de Nacidos Vivos y Defunciones. Available online: https://www.ecuadorencifras.gob.ec/nacimientos-y-defunciones/ (accessed on 3 July 2026).
26. Ministerio de Educación del Ecuador. Registros Administrativos—Archivo Maestro de Instituciones Educativas (AMIE). Available online: https://educacion.gob.ec/base-de-datos/ (accessed on 3 July 2026).
27. Instituto Nacional de Estadística y Censos. VIII Censo de Población y VII de Vivienda. 2022. Available online: https://www.censoecuador.gob.ec (accessed on 3 July 2026).
28. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [CrossRef]
29. Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341. [CrossRef]
30. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.