Article

Not peer-reviewed version

Exploratory Analysis of the Interactions Between Territorial Development Patterns and Electricity Demand in Ecuador's Coastal Region

Diego Peña * , Jorge Murillo , Fernando Ortega , Yadyra Ortiz , Cristian Laverde , Francisco Jurado

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0395.v1

Keywords: electricity demand forecasting; Ecuador; Costa region; principal component analysis; territorial development; multimodel regression; long-horizon projection; energy planning

Article

Exploratory Analysis of the Interactions Between Territorial Development Patterns and Electricity Demand in Ecuador's Coastal Region

Diego Peña [1,2,*], Jorge Murillo [1], Fernando Ortega [3], Yadyra Ortiz [2], Cristian Laverde [1,4] and Francisco Jurado [2]

1 Faculty of Engineering Sciences, Technical State University of Quevedo, Quevedo 120301, Ecuador
2 Department of Electrical Engineering, Universidad de Jaén, 23700 Jaén, Spain
3 Departamento de Ciencias de la Energía y Mecánica, Universidad de las Fuerzas Armadas ESPE, Sangolquí 171103, Ecuador
4 Postgraduate Studies Department, ECOTEC University, Km 13.5, Samborondón, EC092302, Ecuador
* Correspondence: dppb0002@red.ujaen.es

Abstract

This study proposes a reproducible exploratory framework to link long-term territorial development with electricity demand in data-scarce contexts, and applies it to Ecuador's Costa region. The pipeline combines three commonly available input streams: periodic census microdata, an official demand series, and macroeconomic aggregates. Socioeconomic heterogeneity across five non-uniform census rounds (1974, 1982, 1990, 2001, 2010) is summarized through Principal Component Analysis (PCA), and territorial indicators are projected to the demand horizon using low-order polynomial functions. Eleven regression specifications are compared on a log-transformed demand variable, and a rolling-origin backtesting scheme plus a 2020–2024 holdout are used for validation. The selected Trend OLS log model attains $R^2 = 0.551$ and MAPE = 6.08%, and projects a regional demand of approximately 6,940 MW by 2050, equivalent to a compound annual growth rate of 3.45%. Beyond the Ecuadorian case, the results show that transparent, low-data pipelines based on harmonized census information, macroeconomic drivers and simple regression models can provide defensible medium- and long-term demand signals for planners in other emerging economies with limited high-frequency data.

Keywords: electricity demand forecasting; Ecuador; Costa region; principal component analysis; territorial development; multimodel regression; long-horizon projection; energy planning

---

1. Introduction

In developing countries, electricity demand does not respond solely to technical or climatic factors: it is equally shaped by where people live, how the local economy is structured, and by the unequal distribution of infrastructure access [1,2]. This territorial perspective matters because models that ignore the spatial distribution of consumption tend to underestimate imbalances that are later costly to correct. Anticipating future supply requirements therefore demands integrating socioeconomic and spatial dimensions into the analysis, not merely aggregated load data.

Ecuador's Costa region comprising Guayas, Manabí, El Oro, Esmeraldas, Santa Elena, Santo Domingo de los Tsáchilas, and Los Ríos, accounts for approximately 53% of the national population and generates more than 46% of GDP [3,4]. Its warm-humid climate (average temperature 25–28 °C, relative humidity above 70%) produces consumption profiles with high shares of air conditioning and agro-industrial processing [3]. Analysis is further complicated by the region's internal heterogeneity: while Guayas concentrates industrial and commercial activity, Esmeraldas and Los Ríos exhibit higher poverty rates and lower electricity coverage, generating divergent demand structures that aggregate national analyses fail to capture.

 doi:10.20944/preprints202605.0395.v1

The 2023–2024 energy crisis revealed the extent to which these territorial asymmetries carry real operational consequences. The dependence of the National Interconnected System (SNI) on hydroelectric generation which accounted for 67.4% of total production in 2024 made it critically vulnerable to hydrological variability; a prolonged drought resulted in blackouts of up to 14 hours per day [5]. The problem was therefore not solely one of installed capacity: it was also one of mismatch between the spatial distribution of generation and the territorial patterns of demand, a type of structural gap that conventional energy planning approaches frequently underestimate.

The literature on tropical coastal regions documents that temperature accounts for between 40% and 60% of consumption variability ($R^2 = 0.40$–0.60), with typical increases of 2%–4% per degree Celsius attributable primarily to air conditioning use; Pearson coefficients reported range from 0.65 to 0.85 [6–8]. Relative humidity shows moderate association with consumption ($r = 0.35$–0.55) [4,9], while the effects of phenomena such as El Niño on demand in the Costa region have not been rigorously quantified [10–12]. In the socioeconomic dimension, income elasticities lie between 0.6 and 1.2 in developing economies [13], and residential price elasticities range from -0.2 to -0.5 in the short run and from -0.4 to -0.8 in the long run [2,14]. Territorial configuration variables such as population density, urbanization, and household size show significant positive correlations with consumption ($r = 0.55$–0.75), while poverty and limited access to basic services are negatively associated [3,4]. At the temporal level, monthly seasonality accounts for between 15% and 25% of variability in tropical zones, with afternoon peaks that can exceed base demand by up to 40% [15]. Multivariate models integrating these three dimensions achieve explanatory levels of $R^2 = 0.75$–0.90 [2], confirming that no single category of predictors is sufficient on its own.

Machine learning tools have considerably expanded the possibilities for territorial analysis. LSTM networks achieve prediction errors (MAPE) of 1.5%–3.5%, compared to 4%–7% for ARIMA models [16,17]; Random Forest and XGBoost reach MAPE of 2%–4% with $R^2 > 0.95$ [18]; and hybrid models with time series decomposition yield additional precision gains of 15%–25% [19]. However, the transfer of these methods to data-scarce economies faces important constraints, as consumption patterns [20,21], technological access, and energy infrastructure differ substantially from the contexts in which these algorithms were originally developed [22,23]. In Ecuador, recent studies illustrate both the potential and the limitations of this field: Chicaiza Yugcha et al. [12] obtained a MAPE of 3.2% and R^2 of 0.94 at the subnational scale using multiple linear regression; Carrillo Calderón et al. compared Random Forest and XGBoost in Chimborazo with accuracies above 92%; Moya et al. [24] identified seven determinants of residential consumption at 1 km^2 resolution, including population density, GDP, and HDI; Araujo-Vizuete et al. [25] documented significant differences between Quito and Guayaquil associated with income and technological access; Icaza-Alvarado et al. [26] reported temperature–demand correlations between $R^2 = 0.70$ and $R^2 = 0.85$ with income elasticities of 0.8 to 1.1; and Naranjo-Silva et al. [23], using the LEAP model, documented the 2023–2024 crisis and estimated losses of approximately USD 2 billion. Despite these advances, disaggregated analysis of the Costa region remains limited.

The available evidence taken together indicates that climatic variables particularly temperature and humidity significantly influence electricity demand in tropical regions. Socioeconomic and demographic factors such as income, urbanization, population density, and access to infrastructure also strengthen the explanatory power of models when integrated in multivariate approaches. The literature further shows that machine learning methods and hybrid models generally outperform traditional statistical approaches in well-resourced settings. In Ecuador, the reliability problems of the electricity system are closely linked to the territorial distribution of generation and demand. Yet knowledge gaps persist for the Costa region, owing to the scarcity of comprehensive studies and limited availability of consumer-level data.

Against this background, this study analyzes the relationship between territorial development patterns and electricity demand in Ecuador's Costa region. Its contribution lies in jointly integrating

 doi:10.20944/preprints202605.0395.v1

territorial, socioeconomic, and macroeconomic factors within a transparent, reproducible framework to generate evidence useful for more robust, equitable, and resilient energy planning.

Although several studies have addressed electricity demand modeling in Latin America, few frameworks are explicitly designed for contexts in which high-frequency socioeconomic information is unavailable and where the only territorial signal comes from widely spaced national censuses. In Ecuador, for example, the national population and housing censuses used in this work correspond to 1974, 1982, 1990, 2001 and 2010, which defines a non-uniform temporal grid that does not align naturally with the annual demand series. This mismatch between slow-moving territorial drivers and annual electrical consumption has been largely overlooked in previous regional studies [5,23,26].

The present work contributes to this gap by proposing an exploratory, reproducible pipeline that can be read at two levels. At a general level, it provides a methodological template for linking heterogeneous territorial indicators with long-term demand in any region where census microdata, an official demand series and macroeconomic aggregates are available. At an applied level, it instantiates that template on Ecuador's Costa region, where the combination of rapid urbanization, structural transformation and persistent data scarcity makes long-horizon planning particularly challenging [20,22,23].

The remainder of the article is organized as follows. Section 2 describes the materials and the general methodological design, including the census harmonization strategy, the PCA-based synthesis of territorial indicators, the polynomial projection of socioeconomic drivers and the rolling-origin validation protocol. Section 3 reports the empirical results for the Costa region, including model comparison, selected specification and long-term projections. Section 4 discusses the main findings, the conditions under which the framework can be transferred to other emerging economies, and its limitations. Section 5 concludes.

2. Materials and Methods

2.1. Materials

The analytical database was constructed from three complementary official sources. First, census microdata from IPUMS International [27], Censos Ecuador, and the National Institute of Statistics and Censuses (INEC) [28] were used, corresponding to the national censuses of 1974, 1982, 1990, 2001, 2010, and 2022. From these records, 34 sociodemographic variables at the provincial level were derived, covering demographic structure, basic service coverage, educational attainment, employment, and technological penetration. Second, the historical series of annual electricity demand, expressed in MW, was compiled from official CENACE reports for the period 2000–2024 [29]. Third, three macroeconomic indicators from the Central Bank of Ecuador were incorporated: gross domestic product, industrialization index, and total population [30]. Computational analyses were conducted in Python 3.10 and RStudio.

The geographic scope of the study was restricted to the seven provinces of Ecuador's Costa region: Guayas, Manabí, El Oro, Esmeraldas, Santa Elena, Los Ríos, and Santo Domingo de los Tsáchilas. Records corresponding to the Sierra, Amazonia, and Insular regions were excluded to ensure spatial consistency of the analysis. Although Ecuador has census data available through 2022, this study used only the 1974–2010 series, because several of the selected micro-variables underwent changes in their definition, disaggregation level, or collection structure across census rounds. Including the 2022 census would have compromised temporal comparability, as well as the stability of the correlation matrix and the interpretability of the principal components. For this reason, a homogeneous series of five complete census observations was used.

2.2. Methods

2.2.1. General Methodological Design

Before describing the Ecuadorian implementation, it is useful to present the methodological pipeline at a general level, so that its logic can be evaluated independently from the specific case study.

The framework is designed for territorial contexts in which three information layers are typically available: (i) periodic population and housing census microdata, which provide rich but temporally sparse socioeconomic information; (ii) an official annual electricity demand series at the regional or national level; and (iii) macroeconomic aggregates (for example GDP or sectoral value added) published at annual frequency. The pipeline is organized in five sequential stages, summarized in Figure 1 and detailed in Table 1.

In the first stage, census microdata from different rounds are harmonized into a common set of territorial indicators covering demographic structure, household composition, housing quality, access to basic services and educational attainment. In the second stage, Principal Component Analysis is applied to this harmonized matrix to obtain a small number of orthogonal components that summarize the main axes of territorial heterogeneity, reducing multicollinearity and allowing a compact representation of development patterns. In the third stage, each retained component (or each selected indicator) is projected to the demand horizon by means of low-order polynomial functions of calendar year, constrained to degree at most two to avoid spurious curvature given the limited number of census observations. In the fourth stage, the projected territorial drivers are combined with macroeconomic variables and fitted to a log-transformed demand variable through a set of competing regression specifications, which differ in their functional form and in the subset of predictors. In the fifth stage, model selection and validation rely on rolling-origin backtesting, complemented by a final holdout window, so that predictive performance is evaluated under conditions that approximate real planning use.

The five-stage structure is deliberately simple and transparent, which is a design choice rather than a limitation. In data-scarce environments, more complex models tend to overfit and lose interpretability, whereas the present pipeline produces components and coefficients that can be inspected and discussed with domain experts. In methodological terms comparable to Sheng et al., who also relied on PCA-based socioeconomic components [39], and to Boukarta et al., who combined PCA with regression in an urban context [41], the framework emphasizes reproducibility and the ability to be re-estimated whenever new census rounds or updated macroeconomic series become available.

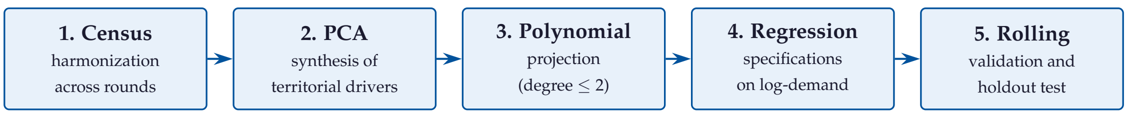


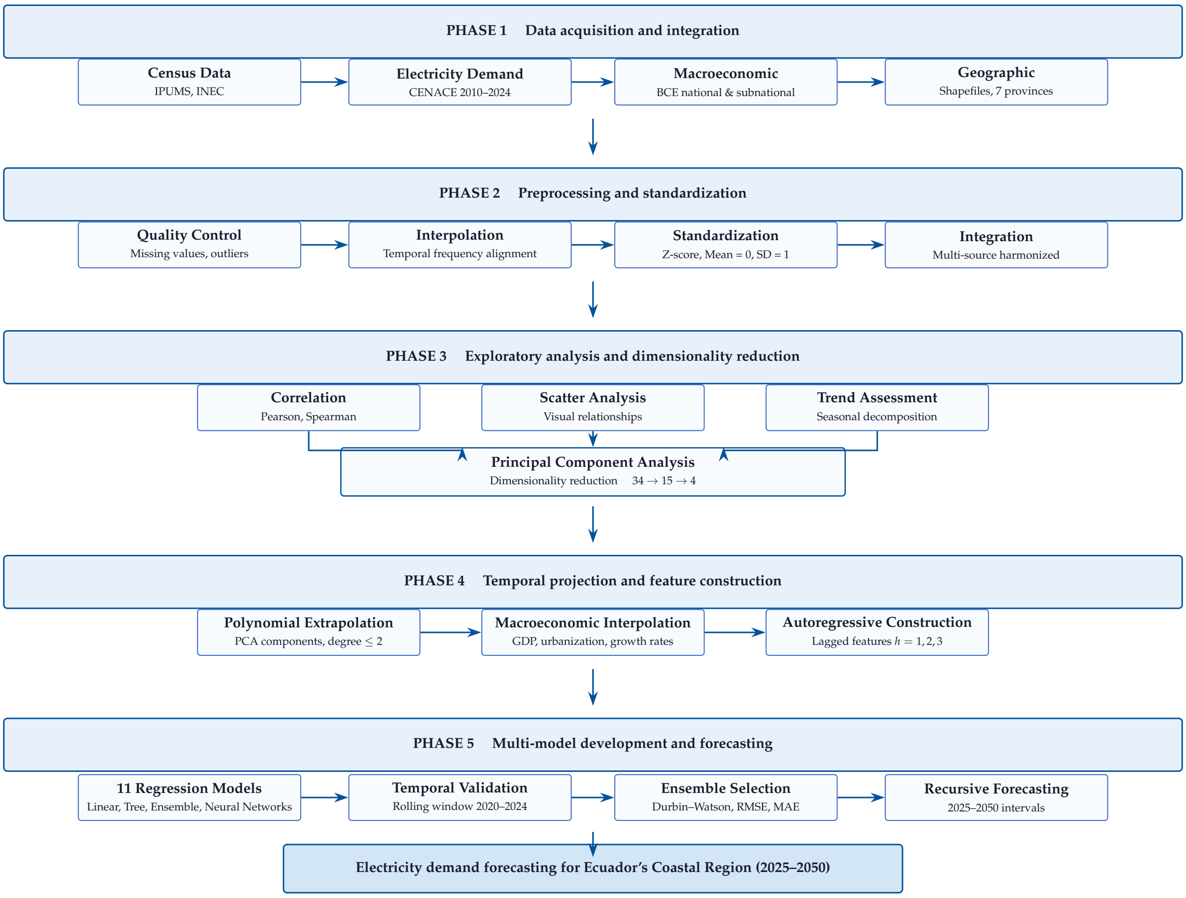
Figure 1. Conceptual five-stage pipeline of the general methodological framework. Its operational instantiation for Ecuador's Costa region is reported in Table 1.

2.2.2. Study Design and Methodological Workflow

Building on the general pipeline introduced above, this subsection describes the specific operational workflow adopted for Ecuador's Costa region. The study adopted a descriptive-correlational quantitative approach to examine the relationships between territorial development patterns and electricity demand in the region. The methodological strategy was structured as a five-phase analytical pipeline integrating dimensionality reduction, temporal projection, and supervised regression techniques. Table 1 summarizes the detailed methodological workflow adopted in the study, which comprises: (i) data acquisition and integration, (ii) preprocessing and standardization, (iii) exploratory analysis and dimensionality reduction, (iv) temporal projection and feature construction, and (v) multi-model development and forecasting.

 doi:10.20944/preprints202605.0395.v1

Table 1. Operational five-phase workflow through which the general methodological framework (Figure 1) is instantiated for Ecuador's Costa region, covering multi-source data integration, preprocessing, dimensionality reduction, feature construction, and electricity demand forecasting.



2.2.3. Data Preprocessing and Standardization

Given the heterogeneity of sources and their irregular periodicity, preprocessing was carried out sequentially in four stages. First, records were cleaned through the identification and removal of outliers and incomplete observations. Second, temporal interpolation was applied to transform the discontinuous census observations into continuous annual series. Third, the selected sociodemographic variables were standardized using Z-score transformation, with zero mean and unit variance. Finally, all sources were integrated into a single analytical dataset with temporal coverage spanning 2000–2050.

2.2.4. Exploratory Analysis and Dimensionality Reduction

The analysis was aimed at identifying distribution and interrelation patterns among the explanatory variables. To this end, the Pearson correlation matrix was computed for the 34 original census variables in order to detect collinearity patterns and guide subsequent predictor selection. In addition, scatter plots were constructed for six pairs of territorial indicators selected for their theoretical relevance to electricity demand, incorporating the corresponding correlation coefficient and an ordinary least-squares trend line. This analysis was complemented by time-evolution plots of demographic and basic service coverage indicators, with the aim of characterizing the socioeconomic transformations underlying electricity consumption patterns.

To mitigate multicollinearity and reduce the predictor space, Principal Component Analysis (PCA) was applied to the subset of variables identified as analytically relevant in the exploratory phase. Four principal components were retained based on cumulative explained variance and the inflection point of the scree plot. The factor loadings and biplot of the first two components were then examined to

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2026 doi:10.20944/preprints202605.0395.v1

interpret the latent structure of the data and verify the temporal consistency of the census observations in the reduced space.

2.2.5. Temporal Projection and Feature Construction

Since the principal component scores were available only at five census points, each retained component was projected to the 2000–2050 horizon using univariate polynomial regression of degree less than or equal to two, selected according to the coefficient of determination. This degree restriction was applied to avoid unstable or divergent behavior during extrapolation. The projected components were subsequently combined with independently interpolated macroeconomic indicators and three lagged variables of the target variable, corresponding to demand at $t-1$, $t-2$, and $t-3$.

2.2.6. Multi-Model Development, Validation, and Forecasting

Eleven regression algorithms grouped into different model families were evaluated. All specifications were estimated on electricity demand transformed using the natural logarithm, in order to stabilize series variance, ensure positive predictions upon back-transformation, and facilitate interpretation of results in terms of relative growth.

Model performance was evaluated using rolling-origin backtesting and a holdout validation period corresponding to 2020–2024. The primary selection criteria prioritized the Durbin–Watson statistic and RMSE. Finally, the three best-performing models were integrated into a simple-average ensemble, and the forecast was executed recursively through 2050.

2.2.7. Formal Specification of the Pipeline

For reproducibility and to facilitate re-estimation of the framework in other emerging economies, the main operational steps are summarized in compact notation. Let y_t denote the annual regional electricity demand (MW) in year t, and let t_0 be the first year of the historical sample.

After standardization of the 15 retained socioeconomic variables, Principal Component Analysis was applied to the pooled census matrix, yielding four components $\mathrm{PC}_i(c)$ scored at each round $c \in \{1974, 1982, 1990, 2001, 2010\}$. Because only five observations are available per component and the forecast horizon extends four decades beyond the last census, each score was extrapolated to 2000–2050 by a univariate linear regression on the calendar year,

$$\widehat{\mathrm{PC}}_i(t) \;=\; \alpha_{i,0} + \alpha_{i,1}\, t, \qquad i = 1, \ldots, 4. \tag{1}$$

A linear trajectory was preferred over higher-degree polynomials because, with only five census points, quadratic extrapolation produces divergent behavior at long horizons.

The forecasting models operate on the log-transformed demand series $\ell_t = \log(y_t)$. The general specification of the pipeline combines a linear trend on the calendar year, the four projected components, and a vector of macroeconomic indicators $\mathbf{m}_t = (\mathrm{GDP}_t, \mathrm{IND}_t, \mathrm{POP}_t)^\top$ (per capita GDP, industrial index, and total population),

$$\ell_t \;=\; \beta_0 + \beta_\tau\,(t - t_0) + \sum_{i=1}^{4} \beta_i\, \widehat{\mathrm{PC}}_i(t) + \gamma^\top \mathbf{m}_t + \varepsilon_t, \tag{2}$$

with ε_t a zero-mean error term. The eleven estimators evaluated share this predictor structure and differ only in their estimation rule: unpenalized least squares, ridge, lasso, ElasticNet, Huber, linear-kernel SVR, polynomial-ridge (degree 2 on the same inputs), spline-ridge (cubic B-spline transformation), and two tree-based nonparametric regressors (random forest and gradient boosting). The autoregressive variant, ARX-ridge, extends Equation (2) with three lags of log-demand,

$$\ell_t \;=\; \beta_0 + \beta_\tau\,(t - t_0) + \sum_{i=1}^{4} \beta_i\, \widehat{\mathrm{PC}}_i(t) + \gamma^\top \mathbf{m}_t + \sum_{k=1}^{3} \phi_k\, \ell_{t-k} + \varepsilon_t, \tag{3}$$

and is the only specification that incorporates past demand explicitly. The recommended model, *Trend OLS log*, corresponds instead to the parsimonious restriction of Equation (2) in which only the trend term is retained,

$$\ell_t \;=\; \beta_0 + \beta_\tau\,(t - t_0) + \varepsilon_t, \tag{4}$$

fitted by ordinary least squares on the 2000–2019 training window and projected recursively to 2050.

Temporal generalization was assessed through rolling-origin backtesting. For each origin t^* within the historical window, the model was re-estimated using the expanding sample $\{(\mathbf{x}_s, \ell_s) : s \le t^*\}$ and used to predict ℓ_{t^*+1}, so that every test point remained strictly out of sample. The 2020–2024 window was reserved entirely as an independent holdout and did not enter any training or hyperparameter-tuning stage.

Predictive accuracy was summarized on the MW scale after back-transforming the predictions as $\hat{y}_t = \exp(\hat{\ell}_t)$, using the root mean squared error, the mean absolute percentage error, and the coefficient of determination,

$$\mathrm{RMSE} \;=\; \sqrt{\frac{1}{n}\sum_{t=1}^{n}\big(y_t - \hat{y}_t\big)^2}, \quad \mathrm{MAPE} \;=\; \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \quad R^2 \;=\; 1 - \frac{\sum_{t=1}^{n}\big(y_t - \hat{y}_t\big)^2}{\sum_{t=1}^{n}\big(y_t - \bar{y}\big)^2}. \tag{5}$$

A positive R^2 on the holdout window therefore indicates that the model outperforms the sample-mean predictor over 2020–2024. The Durbin–Watson statistic was additionally reported as a residual diagnostic (Table 3).

3. Results

3.1. Exploratory Analysis, Dimensionality Reduction, and Socioeconomic Projections

The historical electricity demand series for Ecuador (2000–2024) shows a sustained, nearly monotonic growth trajectory. Demand increased from 1,076.1 MW in 2000 to 2,914.1 MW in 2024, with an average of 1,914.3 MW. Overall, this behavior represents a compound annual growth rate (CAGR) of 3.87% and a cumulative increase of 170.8%, associated with economic expansion, rising consumption, and advances in urbanization and industrialization. The only year-on-year contractions were observed between 2016 and 2017, in line with the macroeconomic adjustment recorded in that period. Also notable is the acceleration during the holdout validation window (2020–2024), whose CAGR reached 7.70% above the historical average and consistent with a stronger post-pandemic recovery.

The census database used to construct structural predictors encompasses 34 socioeconomic variables observed at five census points between 1974 and 2010. These variables were grouped into four thematic domains: (i) household structure and demographics, (ii) access to basic services and infrastructure, (iii) education, employment, and gender composition, and (iv) connectivity and telecommunications. Subsequently, through a sequential filtering and selection process, 15 variables were retained based on their theoretical relevance and low redundancy. The temporal evolution of these variables across the five censuses is presented in Table 2.

 doi:10.20944/preprints202605.0395.v1

Table 2. Evolution of the 15 socioeconomic variables selected for PCA in Ecuador (1974–2010). Proportions are expressed as percentages and Δ indicates the total change between 1974 and 2010.

Code	Socioeconomic Variable	1974	1982	1990	2001	2010	Δ 1974–2010
Domain I — Household structure and demographics ($n = 5$)							
Persons/Household	Persons per household	7.30	6.16	5.40	4.87	4.78	−34.5%
Median Age	Median age (years)	22.73	23.98	24.65	26.18	28.91	+27.2%
Population Density	Population density (inhab/km^2)	54.00	60.42	65.83	72.55	81.88	+51.6%
Rooms/Dwelling	Rooms per dwelling	6.86	6.01	5.41	4.88	4.27	−37.8%
Families/Household	Families per household	2.52	2.39	2.22	2.13	2.08	−17.3%
Domain II — Access to basic services and infrastructure ($n = 4$)							
Electricity Access	Electricity access (%)	47.0	63.2	75.8	89.6	96.0	+104.3%
Water Access	Access to drinking water (%)	18.0	27.5	37.9	51.2	58.3	+224.2%
Kitchen Access	Kitchen access (%)	73.0	75.2	77.1	79.4	80.7	+10.6%
Homeownership Rate	Homeownership rate (%)	63.0	63.4	63.8	64.3	65.6	+4.1%
Domain III — Education, employment, and gender composition ($n = 3$)							
University Education	University education (%)	1.0	1.8	2.9	4.5	6.3	+534.2%
Employment Rate	Employment rate (%)	97.0	96.4	95.8	95.0	93.8	−3.3%
Male Proportion	Male proportion (%)	53.0	52.4	51.8	50.9	50.2	−5.2%
Domain IV — Connectivity and telecommunications ($n = 3$)							
Internet Access	Internet access (%)	0.0	0.0	0.0	2.1	29.0	$+\infty$
Landline Access	Landline telephone (%)	0.0	0.2	0.8	2.2	3.6	$+\infty$
Mobile Phone Access	Mobile phone (%)	0.0	0.0	0.0	3.8	11.6	$+\infty$

The temporal evolution of the selected variables indicates substantial transformations during the 1974–2010 period. In Domain I, the mean number of persons per household decreased by 34.5% (from 7.30 to 4.78), while population density increased by 51.6% (from 54.00 to 81.88 inhab/km^2), evidencing a sustained process of household nucleation and urban expansion. In Domain II, basic service coverage indicators show the most pronounced changes: electricity access nearly doubled (+104.3%, from 47.0% to 96.0%) and access to drinking water tripled (+224.2%, from 18.0% to 58.3%), reflecting the sustained infrastructure investment effort by the Ecuadorian state. In Domain III, the proportion with university education experienced the largest relative increase from 1.0% to 6.3%, albeit from an extremely low base marking the process of mass expansion of higher education. Domain IV (connectivity) started from zero values in 1974 and 1982, with internet penetration reaching 29.0% in 2010 and mobile telephony 11.6%, exhibiting exponential growth dynamics unmatched in any other domain.

Pearson correlation analysis applied to the 34 original socioeconomic variables revealed strong inter-predictor dependence, particularly among demographic, infrastructure, and socioeconomic activity groups. Examining the subset of 15 selected variables yielded 105 possible bivariate pairs, of which 41 (39.0%) exhibited $|r| > 0.80$, 25 (23.8%) exceeded $|r| > 0.90$, and 18 (17.1%) reached $|r| > 0.95$. Taken together, these results confirm high redundancy among variables and justify the application of PCA as a strategy to reduce multicollinearity and synthesize system information (Figure 2).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2026 doi:10.20944/preprints202605.0395.v1

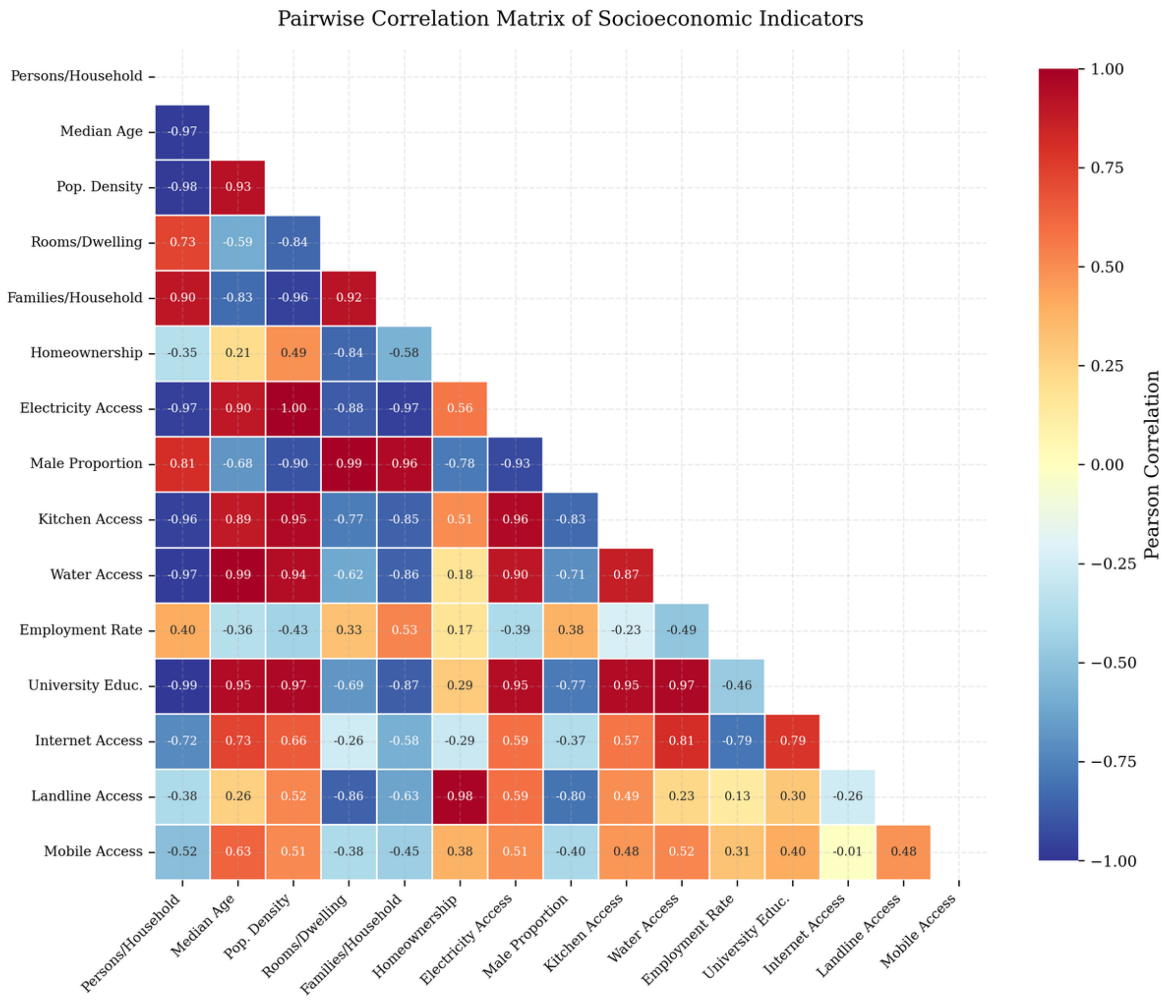


Figure 2. Pearson correlation heatmap for the 15 selected socioeconomic variables.

PCA applied to these 15 variables yielded four principal components (PC1–PC4) with eigenvalues above 0.8. PC1 explained 69.26% of total variance, PC2 19.44%, PC3 8.77%, and PC4 2.53%, with the first two components jointly accounting for 88.70% and all four retained components for 100% of the variance (Figure 3). The loading structure indicates that PC1 is primarily defined by variables linked to socioeconomic modernization, access to basic services, and housing conditions, while PC2 represents a secondary dimension of differentiation associated with demographic, employment, and connectivity variables, as illustrated in the biplot (Figure 4).

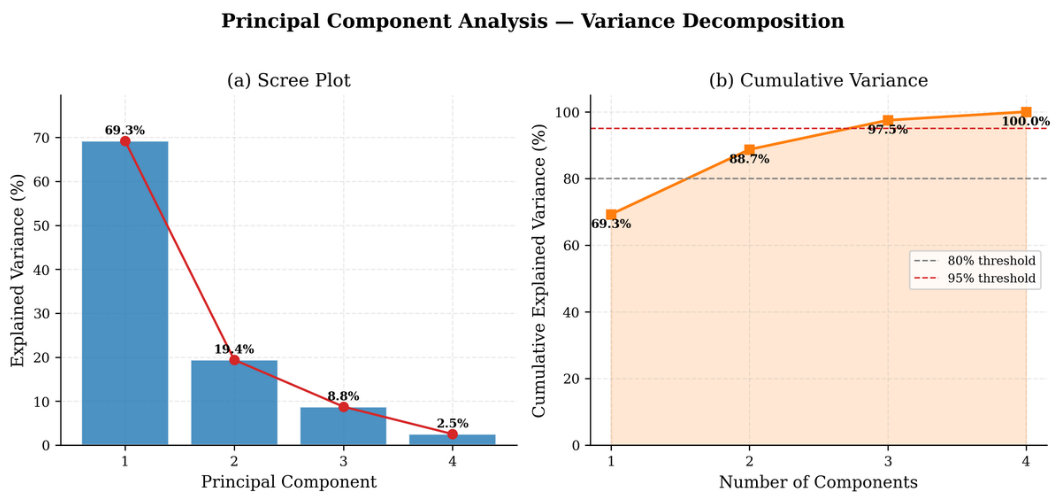


Figure 3. Scree plot and component retention criterion.

 doi:10.20944/preprints202605.0395.v1

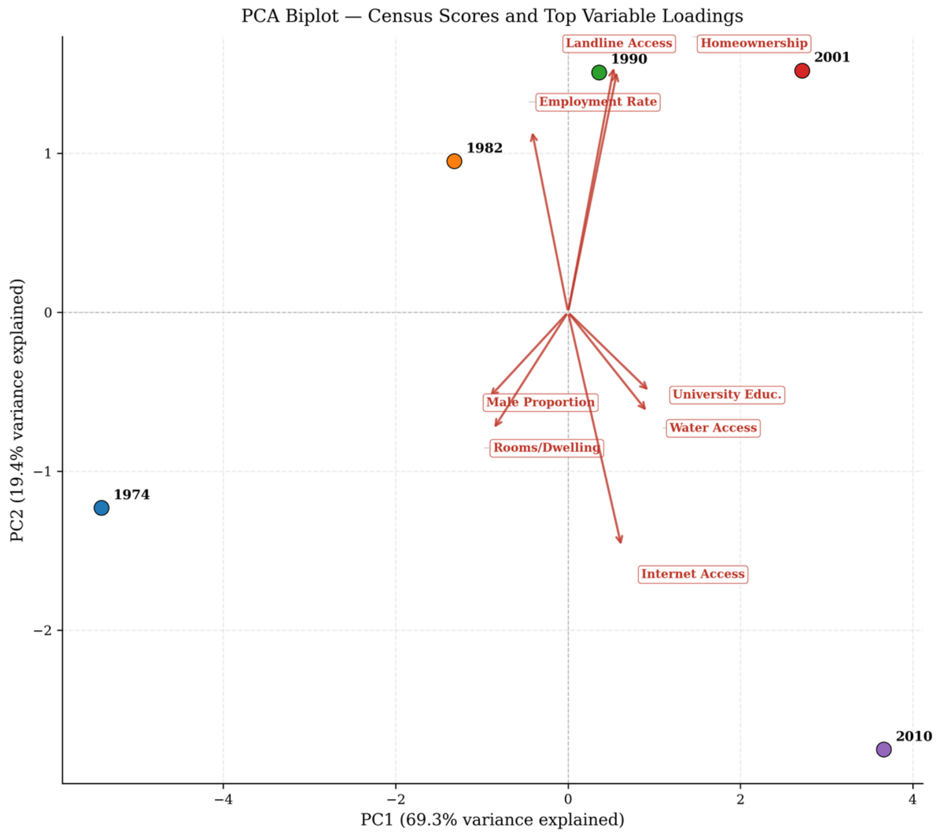


Figure 4. PCA biplot (PC1–PC2) of variables and historical censuses.

The temporal projection of component scores through 2050 was performed via univariate linear regression of each component against the census year. PC1 showed a high fit ($R^2 = 0.929$; slope $= 0.241$ score/year), indicating a consistent temporal trend that supports its extrapolation under a stability criterion. PC2, PC3, and PC4 showed considerably lower fits ($R^2 < 0.06$); their future trajectories were therefore estimated using a conservative approach with slope restriction, in order to limit speculative divergence arising from long-horizon extrapolation with a limited number of observations. These projected trajectories, combined with independently extrapolated macroeconomic variables (per capita GDP, industrial index, and total population), formed the vector of exogenous predictors used across the full set of forecasting models.

3.2. Model Fit During the Training Period

For electricity demand projection, eleven model specifications were implemented, all estimated on demand transformed using the natural logarithm. Several models achieved high in-sample fit during the training phase, but performance was not uniform under temporal validation. Gradient Boosting achieved a near-perfect training fit ($R^2 \approx 1.000$) but showed a substantial drop in backtesting accuracy, a pattern clearly indicative of overfitting. In contrast, regularized models Ridge, Lasso, ElasticNet, and Huber exhibited more stable behavior and a better balance between in-sample fit and out-of-sample predictive capacity, a result consistent with the constraints imposed by the training sample size.

3.2.1. Model Performance in Temporal Backtesting

Rolling-origin backtesting, based on iterative retraining of each model using information available up to each test point within the historical period, allowed evaluation of temporal generalization capacity under a prospective scheme. As shown in Table 3, Gradient Boosting recorded the lowest backtesting RMSE (93.34 MW), followed by Spline Ridge (98.41 MW) and ARX Ridge (109.01 MW). The Durbin Watson statistic showed heterogeneous residual autocorrelation patterns: Trend OLS ($DW = 0.172$) and Poly2 Ridge ($DW = 0.027$) exhibited strong positive autocorrelation, while Gradient

Boosting ($DW = 1.981$), Spline–Ridge ($DW = 1.945$), and Random Forest ($DW = 1.705$) showed more temporally independent residuals.

Table 3. Summary of performance of the eleven models in training, temporal validation, and holdout 2020–2024. RMSE in MW, MAPE in %, DW = Durbin–Watson. (†) Recommended model.

Model	Train R^2	BT RMSE	BT MAPE	BT DW	HO R^2	HO RMSE	HO MAPE	Scenario
Trend OLS log †	0.912	273.85	12.04	0.172	0.551	176.90	6.08	Moderate
Poly2 + Ridge log	0.899	174.40	7.92	0.027	0.502	186.40	6.53	High
SVR Linear log	0.857	118.92	4.11	0.800	0.422	200.85	7.19	High
ARX + Ridge log	0.969	109.01	3.74	0.918	−0.909	364.86	13.79	Conservative
ElasticNet log	0.954	116.01	4.61	1.472	−1.374	406.94	15.65	Moderate
Ridge log	0.966	110.27	4.34	1.440	−2.293	479.19	18.32	Moderate
Lasso log	0.980	132.14	4.73	1.181	−8.211	801.49	30.01	Moderate
Huber log	0.979	119.11	4.41	0.735	−8.730	823.75	30.70	High
Spline + Ridge log	0.999	98.41	4.41	1.945	−1.185	390.36	10.84	Decreasing*
Random Forest	0.981	117.20	4.88	1.705	−1.836	444.71	14.08	Decreasing*
Gradient Boosting	1.000	93.34	4.25	1.981	−2.202	472.60	14.54	Decreasing*

BT = rolling-origin backtesting (2018–2019); HO = temporal holdout (2020–2024). $R^2 < 0$ indicates that the model does not outperform the naive predictor. * Models with a Decreasing scenario (CAGR $<$ 0%) were excluded from the final selection.

3.2.2. Temporal Holdout Validation (2020–2024)

The 2020–2024 period was reserved entirely for holdout evaluation, with no involvement in any training phase or hyperparameter selection. Electricity demand increased from 2,165.6 MW in 2020 to 2,914.1 MW in 2024 a cumulative increase of 34.6% and a CAGR of 7.70%, clearly above the long-run historical trend (3.87%). This increase, partly attributable to post-pandemic economic reactivation and the expansion of rural electrification, constituted a relevant challenge for models calibrated exclusively on pre-2020 information (Figure 5).

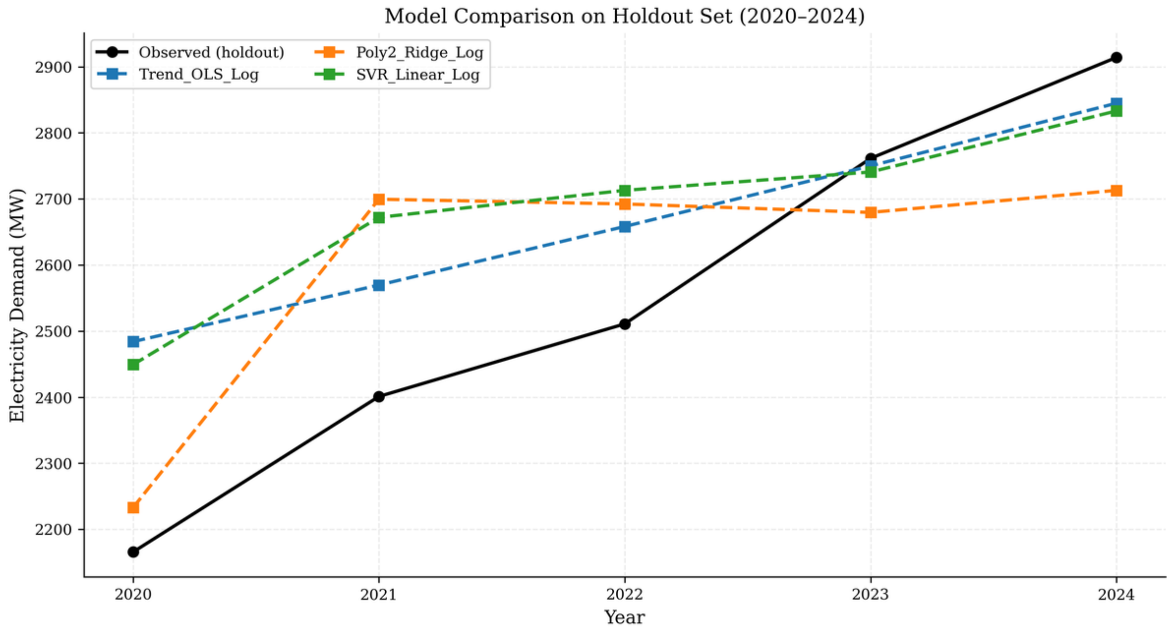


Figure 5. Observed demand and holdout predictions (2020–2024) for the three models with positive R^2.

Under this scenario, only three models achieved a positive coefficient of determination in the holdout set the minimum condition to outperform the naive reference predictor: Trend OLS log ($R^2 = 0.551$, RMSE = 176.9 MW, MAPE = 6.08%), Poly2 Ridge log ($R^2 = 0.502$, RMSE = 186.4 MW, MAPE = 6.53%), and SVR Linear log ($R^2 = 0.422$, RMSE = 200.8 MW, MAPE = 7.19%). The remaining eight models exhibited negative R^2, with RMSE ranging from 364.9 MW to 823.7 MW, reflecting the inability of highly penalized models to adapt to the structural acceleration of post-2020 demand from macroeconomic predictors projected in advance.

Annual analysis of the recommended model (Table 4) shows systematic overestimation in 2020–2022, with errors between +5.9% and +14.7%, attributable to the transient demand decline during the pandemic not anticipated by the macroeconomic predictors. From 2023 onward, the error decreases markedly, reaching −0.4% in 2023 and −2.4% in 2024. This behavior indicates that Trend OLS log consistently reproduces the long-run structural trajectory, although with lower sensitivity to short-term shocks.

Table 4. Holdout predictions (2020–2024) versus actual values.

Year	Actual (MW)	Trend OLS†	Error (%)	ElasticNet	Error (%)	SVR Lin.	Error (%)	ARX Ridge / Error (%)
2020	2,165.6	2,484.0	+14.7	2,449.2	+13.1	2,210.8	+2.1	2,495.9 / +15.3
2021	2,401.0	2,569.6	+7.0	2,946.8	+22.7	2,157.9	−10.1	2,940.6 / +22.5
2022	2,510.9	2,658.3	+5.9	3,001.2	+19.5	2,163.7	−13.8	2,942.1 / +17.2
2023	2,761.4	2,749.9	−0.4	3,026.0	+9.6	2,202.3	−20.2	2,964.0 / +7.3
2024	2,914.1	2,844.8	−2.4	3,113.1	+6.8	2,210.8	−24.1	3,109.8 / +6.7

(†) Recommended model. Negative error values indicate underestimation; positive values indicate overestimation.

3.2.3. Selection Criteria and Recommended Model

The selection of the long-horizon projection model was carried out in two stages. First, models whose projected trajectories for 2025–2050 showed a negative CAGR were discarded, as this behavior is inconsistent with the expected population growth and electrification expansion in Ecuador: Spline–Ridge (CAGR = −0.17%), Random Forest (−0.36%), and Gradient Boosting (−0.02%) were eliminated by this criterion. In a second stage, models with a positive R^2 in the holdout set were prioritized, as this criterion was considered to better reflect predictive capacity in the most recent period. Among the models meeting both requirements, the one with the lowest holdout RMSE was selected.

Under this procedure, Trend OLS log was identified as the most appropriate specification. It combined a coherent prospective trajectory, a positive holdout R^2 (0.551), and the lowest RMSE in the evaluated set (MAPE = 6.08%). Although its log-linear structure is less flexible than other specifications, it provides a stable and monotonic trajectory that reduces the risk of erratic behavior during extrapolation. A methodological limitation should be explicitly acknowledged: the Durbin Watson statistic from backtesting ($DW = 0.172$) indicates positive first-order residual autocorrelation which, while it does not invalidate the central projection, suggests interpreting the results with appropriate caution. Poly2 Ridge log ranked as the second most consistent alternative ($R^2 = 0.502$), albeit with a more expansive projection toward 2050.

3.2.4. Long-Horizon Electricity Demand Forecast (2025–2050)

The selected model projects national electricity demand at 2,925.4 MW for 2025, equivalent to a 0.39% increase over the value observed in 2024 (2,914.1 MW) a continuous transition between the historical series and the projected trajectory, with no abrupt breaks at the start of the forecast horizon. The projection maintains a path of moderate growth, reaching 3,487.2 MW in 2030, 4,955.0 MW in 2040, and 6,940.4 MW in 2050, corresponding to a CAGR of 3.45% for 2024–2050. This pace is lower than that observed in 2020–2024 (7.70%) and slightly below the long-run historical trend (3.87%), suggesting a gradual deceleration toward a more stable growth trajectory (Figure 6).

 doi:10.20944/preprints202605.0395.v1

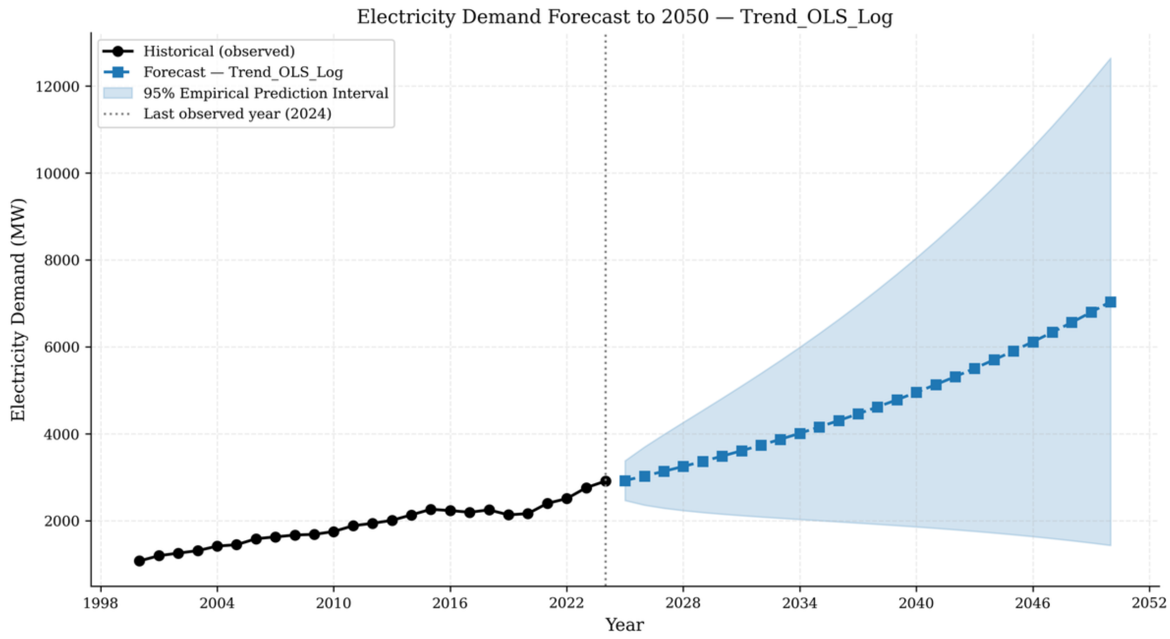


Figure 6. Electricity demand projection 2025–2050 with 95% CI, over the historical series 2000–2024.

Table 5. National electricity demand projections (MW) for 2025–2050 with the recommended model and 95% CI.

Year	Central Projection (MW)	Lower 95% CI (MW)	Upper 95% CI (MW)
2024 (actual)	2,914.1	—	—
2025	2,925.4	2,468.7	3,382.2
2030	3,487.2	2,153.6	4,820.8
2035	4,156.8	2,004.4	6,309.2
2040	4,955.0	1,860.7	8,049.3
2045	5,906.3	1,680.7	10,132.0
2050	6,940.4	1,435.8	12,644.9

The 2024 value corresponds to actual demand. The 95% CI was estimated via residual bootstrap and its width increases with the forecasting horizon.

The confidence interval at 95%, estimated via residual bootstrap, widens significantly as the projection horizon extends. This expansion is inherent in long-horizon forecasts with projected explanatory variables, due to the progressive accumulation of uncertainty at each period. Consequently, results should be interpreted with caution in electricity infrastructure planning applications, and scenario-based planning is recommended over reliance on a single point forecast.

3.2.5. Comparative Scenarios Across All Valid Models

Comparative analysis of the projections generated by all valid models reveals significant divergence across long-horizon scenarios (Table 6). In aggregate, the eight plausible models estimate electricity demand for 2050 ranging from 4,683.4 MW to 9,427.2 MW a spread of 4,744 MW. This dispersion reflects the high uncertainty inherent in long-horizon energy projection exercises, particularly in contexts where historical data availability is limited and the future trajectory depends on structural and macroeconomic variables subject to change. At intermediate horizons, divergence is already significant: for 2040, valid projections range approximately from 3,525 to 5,534 MW, a difference of 2,009 MW (Figure 7).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2026 doi:10.20944/preprints202605.0395.v1

Table 6. Demand projections for 2050, CAGR 2024–2050, and scenario for the eleven models evaluated.

Model	Proj. 2025 (MW)	Proj. 2050 (MW)	Total Growth (%)	CAGR (%)	Scenario
Trend OLS log †	2,925.4	6,940.4	+141.3	3.45	Moderate ✓
Poly2 + Ridge log	2,832.2	9,427.2	+223.3	4.62	High ✓
SVR Linear log	2,979.9	7,372.4	+152.9	3.63	High ✓
ARX + Ridge log	2,505.7	4,683.4	+60.7	1.84	Conservative ✓
ElasticNet log	2,663.5	5,352.3	+83.7	2.37	Moderate ✓
Ridge log	2,834.2	6,426.7	+120.5	3.09	Moderate ✓
Lasso log	2,879.5	6,754.2	+131.7	3.29	Moderate ✓
Huber log	3,000.5	7,918.7	+171.7	3.92	High ✓
Spline + Ridge log	2,961.3	2,790.4	−4.2	−0.17	Decreasing ×
Random Forest	2,609.4	2,654.9	+1.7	−0.36	Decreasing ×
Gradient Boosting	2,888.6	2,898.2	+0.5	−0.02	Decreasing ×

CAGR calculated for 2024–2050. Scenarios: conservative (< 2.0%), moderate (2.0–3.5%), high (3.5–5.0%), and decreasing (< 0%, excluded).

Within this set, the recommended model Trend OLS log projects 6,940.4 MW for 2050 with a CAGR of 3.45%, placing it in an intermediate-to-high position within the range. Its estimate exceeds the more conservative scenarios, such as ARX–Ridge log (4,683.4 MW; 1.84%), ElasticNet log (5,352.3 MW; 2.37%), and Ridge log (6,427.0 MW; 3.10%), but falls below the more expansive trajectories represented by Huber log (7,919.3 MW; 3.89%) and, especially, Poly2–Ridge log (9,427.2 MW; 4.62%). This intermediate position is substantively relevant, as it represents a projection consistent with a moderate growth scenario that avoids both excessive underestimation and overly aggressive demand expansion.

The three models generating decreasing scenarios Random Forest, Gradient Boosting, and Spline Ridge were excluded from the final selection. Their projections for 2050 (2,654–2,898 MW) fall below the value observed in 2024, a result inconsistent with projected population growth, the progressive expansion of electricity coverage, and the electrification dynamics anticipated for Ecuador over the analysis horizon.

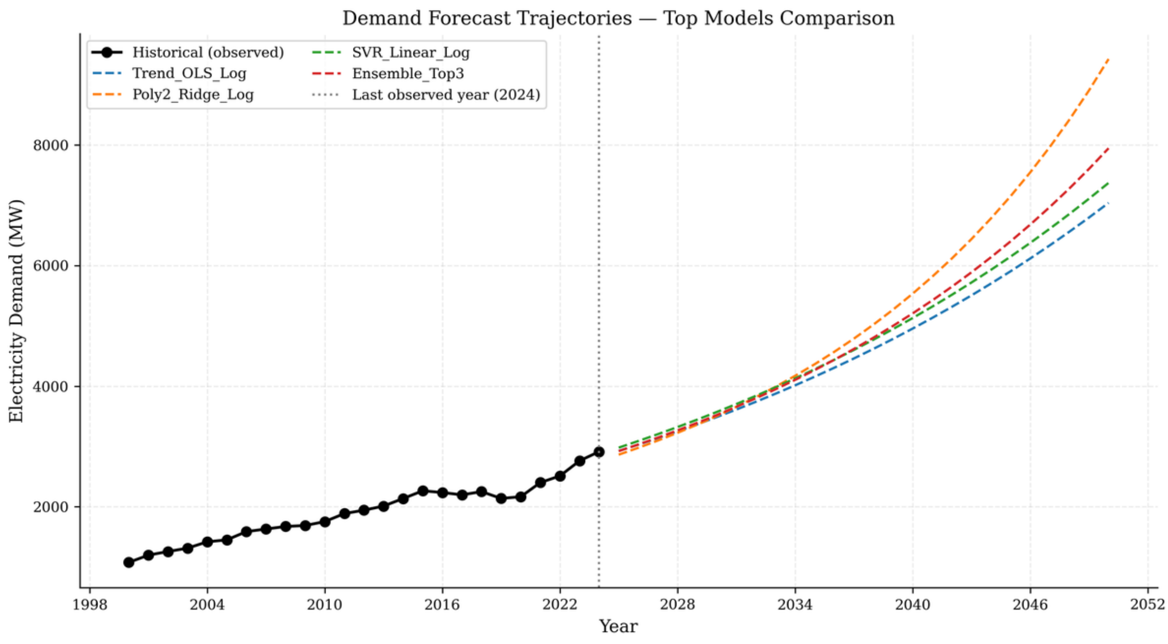


Figure 7. Comparative demand projections 2025–2050 for eight plausible models.

3.2.6. Residual Diagnostics

The residual diagnostics of the recommended model (Trend OLS log) show systematic overestimation during the backtesting period, with a residual mean of −261.8 MW and a standard deviation of 113.6 MW. This indicates that the model tended to project values above those observed, behavior consistent with its trend-based nature in the face of a series that still exhibits cyclical variations not captured by the log-linear specification. In the holdout period (2020–2024), this pattern changes: the initial overestimation decreases gradually until converging toward actual values in 2023, with moderate underestimation appearing in 2024 (Figure 8). Taken together, these patterns suggest that the model did not fully anticipate the post-pandemic demand acceleration, although it did converge toward the observed trajectory in the most recent years.

The Durbin–Watson statistic from backtesting ($DW = 0.172$) confirms the presence of positive first-order residual autocorrelation, indicating that the log-linear formulation does not capture the full temporal dependence of the series. This limitation increases the uncertainty of bootstrap-estimated intervals, so the 95% CI should be interpreted as a conservative planning band rather than a precise probabilistic bound. Residual diagnostics in the holdout period show no systematic long-run deviation after 2022, consistent with the model's gradual convergence toward actual values.

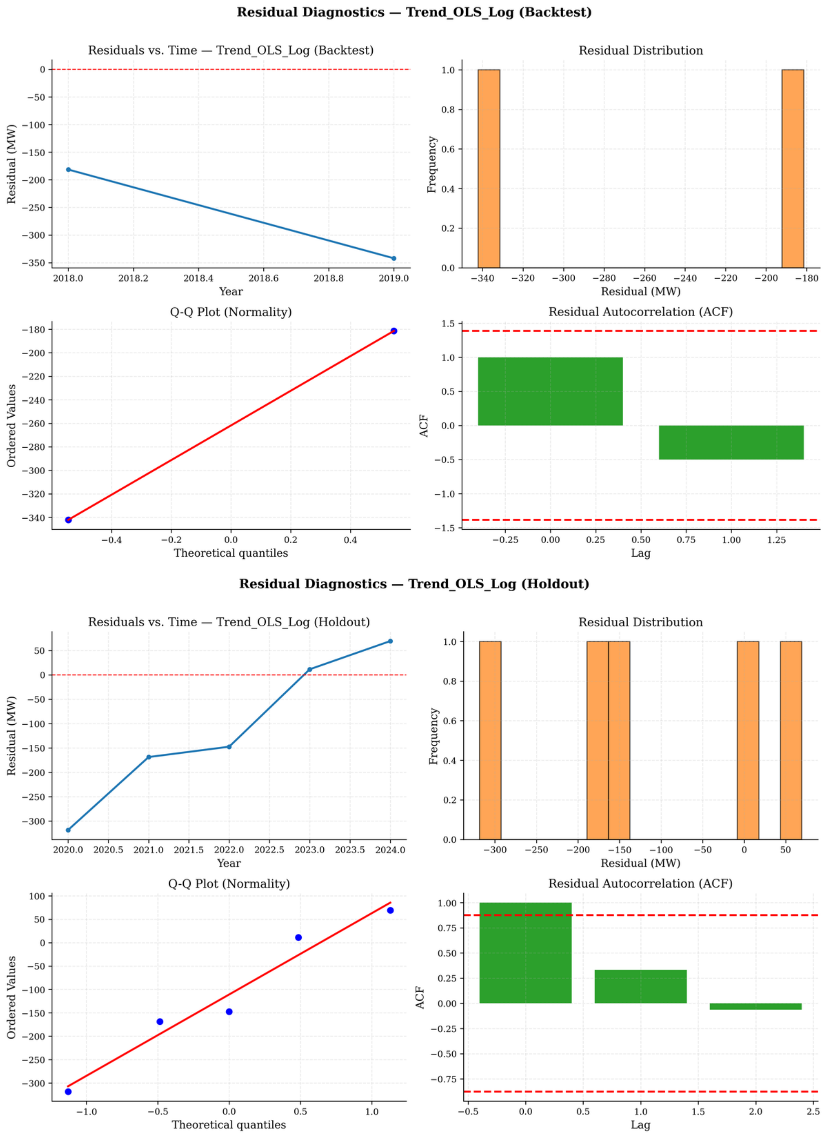


Figure 8. Residual diagnostics of the Trend OLS log model for (a) backtesting and (b) holdout.

4. Discussion

The performance of the recommended model (MAPE = 6.08% in the validation period) should be interpreted in light of data availability. This error level falls within a range comparable to that reported by Mado et al. [31], Vargas-Forero et al. [32], and Hamsa et al. [33] for ARIMA models in contexts with short series. More complex approaches tend to outperform simple specifications when longer series and richer predictor datasets are available [32–36], yet this pattern changes when the sample is small.

 doi:10.20944/preprints202605.0395.v1

Makris et al. [37] show that, in data-scarce scenarios, simpler models can match or even outperform more complex alternatives. This reading aligns with the warning issued by Hippert and Taylor [38] regarding the risk of overfitting and model selection instability in short series. In this context, the fact that a log-linear specification exhibited the most stable behavior in this study should not be interpreted as a methodological limitation, but rather as a coherent response to a problem in which the main constraint is not algorithmic capacity but sample size. In methodological terms comparable to Sheng et al., who also relied on PCA-based socioeconomic components [39], the use of only five census observations as the support for structural predictors limited out-of-sample generalization capacity, even for the most flexible methods. This result is also consistent with the findings of Rúales and Jaramillo et al. [40], who obtained low errors with parsimonious models in Ecuadorian applications.

A second relevant finding is the incorporation of territorial and socioeconomic factors through PCA, in combination with macroeconomic variables and demand lags. Unlike approaches based solely on aggregate indicators, this strategy made it possible to condense a dense and highly interrelated structure linked to income, urbanization, infrastructure access, household size, and connectivity into a small number of components, dimensions widely recognized as determinants of electricity demand [32]. In this regard, the results are consistent with those reported by Sheng et al. [39] and Boukarta et al. [41], who also used PCA to synthesize relevant socioeconomic variables in demand studies. In this study, PC1 and PC2 concentrated the bulk of the variance and provided a consistent reading of the structural modernization process of the Costa region. This finding reinforces the idea that demand expansion depends not only on aggregate macroeconomic changes, but also on long-run territorial and social transformations. In this sense, the results converge with prior work emphasizing the relevance of population density and socioeconomic development in explaining subnational energy consumption [40,43]. Moreover, the internal heterogeneity of the Costa region suggests differentiated trajectories across provinces: while territories such as Guayas, with greater urbanization and industrialization, will likely continue to concentrate a significant share of aggregate demand, other provinces with relatively lower coverage could exhibit faster growth rates as electrification advances [43].

The 2020–2024 validation period also introduced a particularly relevant element: a recent demand acceleration that no model fully captured. The CAGR of 7.70% well above the historical trend suggests an inflection point associated with post-pandemic reactivation, advances in electrification, and the effects of the recent energy crisis [44,45]. This behavior contrasts with simple trend projections that assume continuity in historical patterns, and confirms as already cautioned by Ruales and Jaramillo [40] that atypical events can substantially alter demand dynamics and exceed the explanatory capacity of traditional models. From a planning perspective, this finding reinforces the need to interpret long-horizon projections as plausible trajectories rather than immutable point estimates. Under this logic, the breadth of the projected range for 2050 does not constitute a weakness of the analysis, but a valuable input for designing robust decisions in generation, transmission, and distribution [46,47].

Nevertheless, the results must be interpreted with caution. The structural basis of the model rests on only five census observations, excludes the 2022 census due to comparability issues, works with aggregated regional demand, and does not incorporate climatic variables, despite their relevance in tropical systems [48]. Added to this is the positive residual autocorrelation of the recommended model ($DW = 0.172$), indicating that part of the temporal dependence was not fully absorbed by the specification. This result contrasts with Serrano et al. [49] and Aditya et al. [50], who show that an unresolved residual structure can widen projection intervals and compromise their coverage. For this reason, the projections and their intervals should be treated primarily as a reference for strategic planning, rather than as strict probabilistic bounds. Along these lines, future studies should incorporate climatic variables, deepen the sectoral and territorial disaggregation of demand, and update the structural base with new comparable data sources.

The empirical results for the Costa region should be read jointly with the methodological design presented in Section 2.2.1. The selected Trend OLS log specification, with $R^2 = 0.551$ and MAPE $= 6.08\%$, is not presented as the best possible model in absolute terms, but as the most defensible

 doi:10.20944/preprints202605.0395.v1

specification given the available information. In contexts with richer predictor sets and longer high-frequency series, more sophisticated approaches, such as state-space models or machine-learning regressors, tend to outperform simple log-linear specifications [37,38]. Yet, when the sample is small and the territorial signal comes from widely spaced censuses, parsimonious models with a limited number of interpretable coefficients are often preferable, because they reduce the risk of overfitting and remain auditable by planning authorities.

A second element that deserves attention is the sensitivity of long-term projections to the polynomial extrapolation of territorial drivers. Constraining the polynomial degree to at most two, as described in Section 2, is a conservative choice that prevents implausible curvature at the end of the horizon. Even so, the projected demand of approximately 6,940 MW by 2050 and the implied CAGR of 3.45% should be interpreted as a central trajectory conditional on the continuation of historical development patterns, rather than as a point forecast. This is consistent with the way similar PCA-based exercises have framed their results in other developing contexts [39,41], and with the cautionary notes raised by Arnob and Wang for countries with limited data [20,22].

Finally, the framework contributes to the Ecuadorian planning literature by complementing studies that have focused on reliability, generation mix or scenario-based energy modeling [5,23,40]. Rather than replacing those approaches, the proposed pipeline provides a territorial front-end that can feed any of them, by translating slow-moving census information into synthetic indicators that are comparable across rounds and projectable over time.

4.1. Transferability to Other Emerging Economies

The pipeline described in Section 2.2.1 is intentionally agnostic about the country of application, and its core ingredients are available in many emerging economies. The transferability of the framework, however, is not automatic and depends on several conditions that should be checked on a case-by-case basis. First, at least a few census rounds with comparable variables must be available, so that harmonization and PCA can produce stable components; when only one or two rounds are available, the territorial signal tends to be too weak to support regression-based projection. Second, the demand series should be sufficiently long and free from major structural breaks unrelated to territorial development, such as large tariff reforms or prolonged rationing episodes, which would otherwise dominate the fit. Third, macroeconomic aggregates should be available at annual frequency and at a level of disaggregation that is consistent with the spatial scope of the analysis.

When these conditions are met, the pipeline can be re-estimated with local data without modifying its structure. Countries in the Andean region, Central America and parts of Southeast Asia share with Ecuador a combination of rapid urbanization, structural transformation and limited high-frequency statistics [20,26], which makes the framework potentially useful for their medium- and long-term planning. Where these conditions are only partially met, the same pipeline can still be used as a diagnostic tool, by highlighting which territorial drivers carry most of the explanatory weight and which would require additional data collection before being used for projection.

5. Conclusions

This article proposed and applied a reproducible exploratory framework to analyze the interactions between territorial development patterns and electricity demand in contexts where high-frequency socioeconomic information is scarce. At a general level, the framework combines census harmonization, PCA-based synthesis of territorial indicators, low-order polynomial projection of drivers, multi-model log-linear regression and rolling-origin validation into a single, auditable pipeline. At an applied level, it was instantiated on Ecuador's Costa region using five non-uniform census rounds (1974, 1982, 1990, 2001, 2010) and the official annual demand series.

Among the eleven specifications compared, the Trend OLS log model emerged as the most defensible compromise between fit and parsimony, with $R^2 = 0.551$ and MAPE $= 6.08\%$ on the holdout window. Under the assumption that historical development trajectories continue, the model projects a regional demand of approximately 6,940.4 MW by 2050, equivalent to a compound annual growth rate of 3.45%. This result implies an increase of 141.3% relative to 2024 and confirms the need for sustained expansion of generation, transmission, and distribution infrastructure. These figures should be read as a central planning signal rather than as a deterministic forecast, and the breadth of the scenario range toward 2050 reinforces that planning should rely on a range of plausible trajectories rather than on a single point estimate.

The main contribution of the work is methodological. By making each stage of the pipeline explicit and testable, the framework offers planners and researchers in other emerging economies a transparent template that can be re-estimated whenever new census rounds or updated macroeconomic aggregates become available. Future work will focus on extending the framework to other Ecuadorian regions, on incorporating climatic variables where reliable series exist, on deepening the analysis at the provincial scale, on integrating the 2022 census once its comparability inconsistencies have been resolved, and on comparing the present specification with state-space and machine-learning alternatives once longer high-frequency records are released.

Author Contributions: Conceptualization, D.P.; methodology, D.P.; software, D.P.; validation, D.P., J.M. and F.O.; formal analysis, D.P.; investigation, D.P.; resources, D.P.; data curation, D.P.; writing original draft preparation, D.P. and Y.O.; writing review and editing, D.P., Y.O., C.L. and F.J.; visualization, D.P.; supervision, F.J.; project administration, D.P.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The article processing charge (APC) was waived by the publisher.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the reported results are available upon reasonable request to the corresponding author. Census microdata were obtained from IPUMS International (https://international.ipums.org) and INEC (https://www.ecuadorencifras.gob.ec). Electricity demand data are available in CENACE annual reports (https://www.cenace.gob.ec).

Acknowledgments: The authors would like to thank IPUMS International, INEC, CENACE, and the Central Bank of Ecuador for providing the data used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principal Component Analysis
BCE	Central Bank of Ecuador
CAGR	Compound Annual Growth Rate
DW	Durbin–Watson statistic
GDP	Gross Domestic Product
HDI	Human Development Index
HO	Holdout
LEAP	Long-range Energy Alternatives Planning
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
SNI	National Interconnected System
SVR	Support Vector Regression

doi:10.20944/preprints202605.0395.v1

References

1. Peplinski, M.; Dilkina, B.; Chen, M.; Silva, S.J.; Ban-Weiss, G.A.; Sanders, K.T. A Machine Learning Framework to Estimate Residential Electricity Demand Based on Smart Meter Electricity, Climate, Building Characteristics, and Socioeconomic Datasets. *Appl. Energy* **2024**, *357*, 122413. *doi*: 10.1016/j.apenergy.2023.122413
2. Tang, W.; Wang, H.; Lee, X.-L.; Yang, H.-T. Machine Learning Approach to Uncovering Residential Energy Consumption Patterns Based on Socioeconomic and Smart Meter Data. *Energy* **2022**, *240*, 122500. *doi*: 10.1016/j.energy.2021.122500
3. Carrillo C., D.B.; Pilatuña P., W.P.; Chicaiza Q., C.I. Forecasting Energy Consumption in the Chimborazo Province, Ecuador, Using Random Forest and XGBoost Algorithms. In Proceedings of the 2023 1st International Conference on Advanced Engineering and Technologies (ICONNIC), IEEE, October 2023; pp. 66–72. *doi*: 10.1109/ICONNIC59854.2023.10467276
4. Parra-Jácome, R.M.; Yánez-Jácome, G.B.; Pinto-Arteaga, G.R.; Rea-Toapanta, A.R. Consumos Heterogéneos de Energía en las Tipologías de Hogares del Sector Residencial del Ecuador. *FIGEMPA: Investigación y Desarrollo* **2024**, *17*, 102–111. *doi*: 10.29166/revfig.v17i1.6104
5. Peña, D.; Téllez, A.A.; Jurado, F. Reliability Assessment of Ecuador's Power System: Metrics, Vulnerabilities, and Strategic Perspectives. *Energies* **2025**, *18*, 3059. *doi*: 10.3390/en18123059
6. Cheng, L.; Zang, H.; Xu, Y.; Wei, Z.; Sun, G. Probabilistic Residential Load Forecasting Based on Micrometeorological Data and Customer Consumption Pattern. *IEEE Trans. Power Syst.* **2021**, *36*, 3762–3775. *doi*: 10.1109/TPWRS.2021.3051684
7. Clements, A.E.; Hurn, A.S.; Li, Z. Forecasting Day-Ahead Electricity Load Using a Multiple Equation Time Series Approach. *Eur. J. Oper. Res.* **2016**, *251*, 522–530. *doi*: 10.1016/j.ejor.2015.12.030
8. Rawal, K.; Ahmad, A. Feature Selection for Electrical Demand Forecasting and Analysis of Pearson Coefficient. In Proceedings of the 2021 IEEE 4th International Electrical and Energy Conference (CIEEC), IEEE, May 2021; pp. 1–6. *doi*: 10.1109/CIEEC50170.2021.9510614
9. Chen, P.-C.; Kezunovic, M. Load Consumption Prediction Utilizing Historical Weather Data and Climate Change Projections. In Proceedings of the 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP), IEEE, September 2017; pp. 1–6. *doi*: 10.1109/ISAP.2017.8071415
10. Jawad, M.; Nadeem, M.S.A.; Shim, S.-O.; Khan, I.R.; Shaheen, A.; Habib, N.; Hussain, L.; Aziz, W. Machine Learning Based Cost Effective Electricity Load Forecasting Model Using Correlated Meteorological Parameters. *IEEE Access* **2020**, *8*, 146847–146864. *doi*: 10.1109/ACCESS.2020.3014086
11. Salas-Monteros, J.M.; Maldonado-Navarro, J.L.; Llerena-Poveda, V. del C.; Alban-Navarro, S.F. Evolución del Consumo y Generación de Energía Eléctrica en Ecuador: Análisis del Balance Energético y Diversificación de la Matriz Energética (2021–2024). *Rev. Investig. Talentos* **2025**, *12*, 1–16. *doi*: 10.33789/talentos.12.1.204
12. Chicaiza-Yugcha, O.F.; Martínez-Guaman, C.J.; Orozco-Manobanda, I.A.; Arellano-Castro, Á.D. Previsión del Consumo Eléctrico en el Cantón Salcedo Mediante Técnicas de Aprendizaje Automático. *REVISTA ODIGOS* **2024**, *5*, 9–24. *doi*: 10.35290/ro.v5n1.2024.1134
13. Sun, H.; Han, B. Regional Power Grid Load-Forecast Considering Socio-Economic Factors. In Proceedings of the 2023 2nd International Conference on Advanced Electronics, Electrical and Green Energy (AEEGE), IEEE, May 2023; pp. 70–74.
14. Karuparthi, Y.; Shoaib, S.H.; Kommareddi, H.C.; Jothi, J.A.A. Predicting Energy Demand with Interpretable Machine Learning Models. In Proceedings of the 2024 International Conference on Artificial Intelligence, Metaverse and Cybersecurity (ICAMAC), IEEE, October 2024; pp. 1–6.
15. Jasiński, T. A New Approach to Modeling Cycles with Summer and Winter Demand Peaks as Input Variables for Deep Neural Networks. *Renew. Sustain. Energy Rev.* **2022**, *159*, 112217. *doi*: 10.1016/j.rser.2022.112217
16. Shiwakoti, R.K.; Charoenlarpnopparut, C.; Chapagain, K. A Deep Learning Approach for Short-Term Electricity Demand Forecasting: Analysis of Thailand Data. *Appl. Sci.* **2024**, *14*, 3971. *doi*: 10.3390/app14103971
17. Yajure-Ramírez, C.A. Pronóstico de Consumo de Energía Eléctrica Residencial de Corto Plazo Utilizando Algoritmos de Aprendizaje Automático y Profundo. *Rev. Investig. Sist. Informát.* **2022**, *15*. *doi*: 10.15381/risi.v15i2.23909
18. Abiodun, A.A.; Tosho, A.; Kamil, S.K. Prediction Model for Electricity Energy Consumption and Pricing Using XGBoost with L1 Regularization. *Kasu J. Comput. Sci.* **2024**, *1*, 640–652. *doi*: 10.47514/kjcs/2024.1.3.0017
19. Safari, A.; Kharrati, H.; Rahimi, A. VoltaVistaMan: Energy Dynamics Intelligent Predictive Analysis. In Proceedings of the 2024 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, June 2024; pp. 316–322. *doi*: 10.1109/ICPHM61352.2024.10626498

20. Arnob, S.S.; Arefin, A.I.M.S.; Saber, A.Y.; Mamun, K.A. Energy Demand Forecasting and Optimizing Electric Systems for Developing Countries. *IEEE Access* **2023**, *11*, 39751–39775. *doi*: 10.1109/ACCESS.2023.3250110
21. Hassan, H.; Xiaoying, W.; Sampene, A.K.; Xu, L. The Socio-Economic and Technological Dimensions of Energy Transition. *Energy Strategy Rev.* **2025**, *62*, 101895. *doi*: 10.1016/j.esr.2025.101895
22. Wang, B.; Fu, Q.; Lu, Y.; Liu, K. Limited Data Availability in Building Energy Consumption Prediction. *Information* **2025**, *16*, 575. *doi*: 10.3390/info16070575
23. Naranjo-Silva, S.; Punina-Guerrero, D.J.; Jacome-Dominguez, E.A.; Escobar-Segovia, K.; Laverde-Albarracín, C. Energy Planning Under Climate Pressure in Ecuador: Insights from the 2023–2024 Crisis Using LEAP Modeling. *Sustainability* **2026**, *18*, 2112. *doi*: 10.3390/su18042112
24. Moya, D.; Copara, D.; Borja, A.; Pérez, C.; Kaparaju, P.; Pérez-Navarro, Á.; Giarola, S.; Hawkes, A. Geospatial and Temporal Estimation of Climatic, End-Use Demands, and Socioeconomic Drivers of Energy Consumption in the Residential Sector in Ecuador. *Energy Convers. Manag.* **2022**, *261*, 115629. *doi*: 10.1016/j.enconman.2022.115629
25. Araujo-Vizuete, G.; Robalino-López, A.; Mena-Nieto, Á. Looking beyond Subsidies: Understanding the Complexity of Household Energy Consumption Dynamics of Ecuador's Main Cities. *Cities* **2025**, *163*, 106008. *doi*: 10.1016/j.cities.2025.106008
26. Icaza-Alvarez, D.; Jurado, F.; Tostado-Véliz, M. Smart Energy Planning for the Decarbonization of Latin America and the Caribbean in 2050. *Energy Rep.* **2024**, *11*, 6160–6185. *doi*: 10.1016/j.egyr.2024.05.067
27. IPUMS International. Citation of IPUMS International. Available online: https://international.ipums.org/international/citation.shtml (accessed on 8 April 2026).
28. Instituto Nacional de Estadística y Censos. *Resultados del Censo Nacional de Población*; INEC: Quito, Ecuador, 2022; pp. 1–62.
29. Medina, J. *Operador Nacional de Electricidad – CENACE*; CENACE: Quito, Ecuador, 2024; pp. 1–220.
30. Banco Mundial. PIB per Cápita (UMN Actual). Available online: https://datos.bancomundial.org/indicador/NY.GDP.PCAP.CN (accessed on 8 April 2026).
31. Mado, I.; Rajagukguk, A.; Triwiyatno, A.; Fadllullah, A. Short-Term Electricity Load Forecasting Model Based DSARIMA. *Int. J. Electr. Energy Power Syst. Eng.* **2022**, *5*, 6–11. *doi*: 10.31258/ijeepse.5.1.6-11
32. Vargas-Forero, V.M.; Manotas-Duque, D.F.; Trujillo, L. Comparative Study of Forecasting Methods to Predict the Energy Demand for the Market of Colombia. *Int. J. Energy Econ. Policy* **2024**, *15*, 65–76. *doi*: 10.32479/ijeep.17528
33. Hamsa, H.; Asem, A.; El-Bakry, H. Advanced Time Series Forecasting Models for Electricity Demand Prediction: A Comparative Study. *Fusion: Practice Appl.* **2024**, *15*, 19–31. *doi*: 10.54216/FPA.150102
34. Ko, D.; Yoon, Y.; Kim, J.; Choi, H. Effective Electricity Demand Prediction via Deep Learning. *IEIE Trans. Smart Process. Comput.* **2021**, *10*, 483–489. *doi*: 10.5573/IEIESPC.2021.10.6.483
35. Saha, E.; Saha, R.; Mridha, K. Short-Term Electricity Consumption Forecasting: Time-Series Approaches. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), IEEE, October 2022; pp. 1–5. *doi*: 10.1109/ICRITO56286.2022.9964624
36. Mardotillah, N.A.; Hasanah, R.N.; Wijono. DE-Optimized Hybrid ARIMA-LSTM for Long Term Electricity Load Forecasting. *J. EECCIS (Electrics, Electronics, Communications, Controls, Informatics, Systems)* **2025**, *19*, 39–45. *doi*: 10.21776/jeeccis.v19i2.1813
37. Makris, I.; Moschos, N.; Radoglou-Grammatikis, P.; Andriopoulos, N.; Tzanis, N.; Malamaki, K.-N.; Fotopoulou, M.; Kaousias, K.; Tompros, G.L.; Sarigiannidis, P. Exploring Load Forecasting: Bridging Statistical Methods to Deep Learning Techniques. In Proceedings of the 2024 20th International Conference on Distributed Computing in Smart Systems (DCOSS-IoT), IEEE, April 2024; pp. 491–496.
38. Hippert, H.S.; Taylor, J.W. An Evaluation of Bayesian Techniques for Controlling Model Complexity and Selecting Inputs in a Neural Network for Short-Term Load Forecasting. *Neural Netw.* **2010**, *23*, 386–395. *doi*: 10.1016/j.neunet.2009.11.016
39. Sheng, Y.; Liu, J.; Wei, D.; Song, X. Heterogeneous Study of Multiple Disturbance Factors Outside Residential Electricity Consumption: A Case Study of Beijing. *Sustainability* **2021**, *13*, 3335. *doi*: 10.3390/su13063335
40. Ruales, J.; Jaramillo, M. Medium-Term Electric Network Planning Using SARIMA: EERSA Case Study – Ecuador. In Proceedings of the 2024 11th International Conference on Power and Energy Systems Engineering (CPESE), IEEE, September 2024; pp. 27–32. *doi*: 10.1109/CPESE62584.2024.10841168
41. Boukarta, S.; Berezowska-Azzag, E. Assessing Households' Gas and Electricity Consumption: A Case Study of Djelfa, Algeria. *Quaest. Geogr.* **2018**, *37*, 111–129. *doi*: 10.2478/quageo-2018-0034

42. Moya, D.; Arroba, C.; Castro, C.; Pérez, C.; Copara, D.; Borja, A.; Giarola, S.; Hawkes, A. Long-Term Sustainable Energy Transition of Ecuador's Residential Sector. 2024, pp. 23–40.
43. Cevallos, B.; Urquizo, J. Spatial National Multi-Period Long-Term Energy and Carbon Planning Scenarios in Ecuador's Electric System. *J. Environ. Manage.* **2024**, *370*, 122010. *doi*: 10.1016/j.jenvman.2024.122010
44. Abdurohman, M.; Putrada, A.G. Forecasting Model for Lighting Electricity Load with a Limited Dataset Using XGBoost. *Kinetik: Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control* **2023**, *8*, 571–580. *doi*: 10.22219/kinetik.v8i2.1687
45. Galván, A.; Haas, J.; Moreno-Leiva, S.; Osorio-Aravena, J.C.; Nowak, W.; Palma-Benke, R.; Breyer, C. Exporting Sunshine: Planning South America's Electricity Transition with Green Hydrogen. *Appl. Energy* **2022**, *325*, 119569. *doi*: 10.1016/j.apenergy.2022.119569
46. Rodriguez, A.M.B.; Trotter, I.M. Climate Change Scenarios for Paraguayan Power Demand 2017–2050. *Clim. Change* **2019**, *156*, 425–445. *doi*: 10.1007/s10584-019-02470-1
47. Fields, N.; Collier, W.; Kiley, F.; Caulker, D.; Blyth, W.; Howells, M.; Brown, E. Long-Term Forecasting: A MAED Application for Sierra Leone's Electricity Demand (2023–2050). *Energies* **2024**, *17*, 2878. *doi*: 10.3390/en17122878
48. Trotter, I.M.; Bolkesjø, T.F.; Féres, J.G.; Hollanda, L. Climate Change and Electricity Demand in Brazil: A Stochastic Approach. *Energy* **2016**, *102*, 596–604. *doi*: 10.1016/j.energy.2016.02.120
49. Serrano, A.L.M.; Rodrigues, G.A.P.; Martins, P.H.S.; Saiki, G.M.; Filho, G.P.R.; Gonçalves, V.P.; Albuquerque, R.O. Statistical Comparison of Time Series Models for Forecasting Brazilian Monthly Energy Demand. *Appl. Sci.* **2024**, *14*, 5846. *doi*: 10.3390/app14135846
50. Thangjam, A.; Jaipuria, S.; Dadabada, P.K. Model Selection for Long-Term Load Forecasting under Uncertainty. *J. Model. Manag.* **2024**, *19*, 2227–2247. *doi*: 10.1108/JM2-09-2023-0211