Comparative Analysis of Artificial Neural Networks and Multiple Linear Regression in Electrical Load Forecasting under Limited Data Conditions: A Case Study at Cu Chi Power Company

*Quang Vinh Luu[1], and Quang Minh Huynh[1*1] Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam.*
*Corresponding author: hqminh@hcmut.edu.vn[1*1], lqvinh.sdh231@hcmut.edu.vn[1]*

Abstract: Medium-term electrical load forecasting plays a pivotal role in planning investment, upgrading, and operating distribution grids in urbanizing areas. This study aims to evaluate and quantify the effectiveness of traditional statistical mathematical models versus modern machine learning algorithms in the context of limited planning data at the local level. Based on a 16-year historical dataset (2010 - 2025) in Cu Chi district, three methods including multiple linear regression (OLS), time series (ARIMA), and artificial neural networks (ANN) were es Table III lished to interpolate and extrapolate forecasting scenarios for the 2026-2030 period. Experimental results indicate that, with the characteristic of a small sample size (N=16), the machine learning algorithm suffers from severe overfitting, causing the future forecasting trajectory to be illogically saturated, despite a very low training error of 1.24%. In contrast, the multiple linear regression method maintains high reliability with a mean absolute percentage error of 4.12%, and the future interpolation results are completely consistent with the macroeconomic growth momentum. Consequently, the multiple linear regression method is proven to be the most explicit and reliable tool for load forecasting, directly supporting sustainable energy infrastructure investment decisions under limited data conditions.

Keywords: Electrical load forecasting; Multiple linear regression; Artificial neural network; Overfitting; Energy infrastructure.

I. INTRODUCTION

In the context of robust industrialization and urbanization, energy consumption demand is increasing rapidly [1], imposing immense pressure on power infrastructure systems. Medium-term electrical load forecasting plays a key role in the planning and development of the distribution grid according to current regulations on load research [2]. Formulating accurate forecasting scenarios not only ensures energy security and maintains supply-demand balance but also helps operational management avoid the risks of capital waste due to over-forecasting or power shortages caused by forecasting lower than actual growth. For the local power management level, quantitative forecasting results are a mandatory input parameter to develop plans for renovating and upgrading substations and the medium-voltage distribution system.

To address the load forecasting problem, modern data science has developed a variety of tools, ranging from traditional statistical mathematical models [3],[4] to advanced machine learning algorithms [5],[6]. In recent years, Artificial Neural Networks (ANN) have attracted significant attention due to their superior processing capabilities for complex nonlinear data series. However, the implementation of Artificial Intelligence algorithms in practical grassroots energy management faces two major technical barriers. First, machine learning models often operate as "black boxes," lacking explainability in interpreting causes and quantifying the contribution of each macroeconomic factor to the forecast outcomes. Second, these models require a massive training dataset (Big Data) for the algorithms to accurately converge and optimize their weights [7].

The reality at local power companies reveals that socioeconomic driving indicators such as mechanical population growth, number of enterprises, or revenue scale are typically updated only on an annual basis. This sparse statistical frequency severely limits the size of historical data samples. The forced application of complex machine learning algorithms to small datasets triggers a significant risk of overfitting, rendering the extrapolated future outcomes flattened, illogical, and unreliable. Meanwhile, most current forecasting methods at these units heavily rely on qualitative expert experience or simple trend functions, failing to quantitatively establish the causal correlation between the expansion of the local economy and electrical consumption.

To ensure methodological soundness and meet practical planning needs, this paper conducts a comprehensive comparative analysis of forecasting capabilities between Applied Mathematics (represented by Multiple Linear Regression - MLR/OLS and Time Series - ARIMA) and Artificial Intelligence (represented by Artificial Neural Networks - ANN). The study establishes experiments on a 16-year historical dataset of Cu Chi Power Company, an area experiencing rapid industrial infrastructure development at the northwestern gateway of Ho Chi Minh City.

By evaluating training errors, analyzing future extrapolation trajectories, and examining the degree of causal explanations, the study aims to demonstrate the superiority of the multiple linear regression method under conditions of limited sample sizes. Consequently, the paper proposes a scientific, explicit, and stable quantitative solution to directly support sustainable energy infrastructure investment decisions at grassroots power utilities.

II. RESEARCH METHODOLOGY

The methodological framework of the study is designed based on establishing and comparing three forecasting models representing two schools of thought: Statistical Mathematics (Linear Regression, Time Series) and Machine Learning (Artificial Neural Networks).

A. Multiple Linear Regression Method (MLR/OLS)

Multiple linear regression is a widely applied mathematical method to quantify the causal relationship between a dependent variable (commercial electricity output) and multiple macroeconomic independent variables (population, revenue, and enterprise scale) [8].

In the practice of macroeconomic data collection, these variables often exhibit significant discrepancies in measurement scales and tend to grow simultaneously over time, posing a high risk of multicollinearity. Therefore, to narrow the variance, homogenize the reference frame, and control multicollinearity risks, all input data are pre-processed through natural logarithmic transformation (ln). The proposed general regression equation takes the form:

$$\ln(E_t) = \beta_0 + \beta_1 \ln(Pop_t) + \beta_2 \ln(Doanhthu_t) + \beta_3 \ln(SoDN_t) + \epsilon_t$$

Trong đó:

- E_t: represents electricity consumption.
- $Popt_t$: population.
- $Doanhthu_t$: revenue.
- $SoDN_t$: number of enterprises.
- B_0: intercept.
- B_1,β_2,β_3: parameters are individual slope coefficients reflecting the impact level of each independent variable.
- ϵ_t: random error,

To determine the parameter set β, the Ordinary Least Squares (OLS) algorithm is utilized to find the estimated values such that the sum of the squared residuals is minimized. The post-estimation fitness of the MLR/OLS model is rigorously tested through the coefficient of determination R^2:

$$R^2 = 1 - \frac{RSS}{TSS}$$

The **Adjusted** R^2: $R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$

(to penalize unnecessary variables), t-test significance, and Variance Inflation Factor (VIF). [9]

B. Time Series - ARIMA

To establish a mathematical baseline for comparison with the OLS method, the Autoregressive Integrated Moving Average (ARIMA) model is employed to represent a pure extrapolation approach. Unlike MLR, the ARIMA method tackles the forecasting problem by directly exploiting the regularity and historical inertia of the data series without relying on exogenous variables [10]. The algorithm automatically analyzes and decomposes the historical electrical load data series into trend and random noise components, thereby leveraging the inertia of the time series to shape future forecasting trajectories.

C. Artificial Neural Network - ANN

Representing the artificial intelligence technique group, the Multilayer Perceptron (MLP) architecture is applied to verify the capability to identify and process complex nonlinear relationships between the economy and electrical energy.

To ensure fairness in the comparative experiment, the neural network is designed with an input layer containing exactly three exogenous variables similar to the OLS method (Population, Revenue, Number of enterprises). During the network training process, the data pass through a hidden layer utilizing a Hyperbolic tangent (Tanh) activation function. Based on the backpropagation principle, the algorithm automatically learns and adjusts the linking weights between neurons to minimize forecasting errors on the historical dataset before proceeding to interpolate the future scenario.

D. Evaluation and Comparison Criteria

To objectively evaluate and quantify the accuracy among the three models, the study utilizes two basic statistical indicators: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

Where RMSE is used to evaluate the absolute magnitude of the error (in kWh), heavily penalizing abnormal errors. On the other hand, MAPE is the most crucial and core indicator, expressing the error as a percentage (%), providing the most intuitive and easily interpretable basis to support decision-making in power grid operations [11].

III. CASE STUDY AND DATA

A. Introduction to the Study Area

The comparative experiment of forecasting models is conducted based on actual operational data in Cu Chi District, the northwestern gateway area of Ho Chi Minh City. This area is experiencing rapid urbanization with a population reaching 524,495 people in 2025 and an average GRDP growth rate of 11.8% during the 2010-2025 period. This region concentrates four large-scale industrial parks. The continuous expansion of production and business facilities here is the main driver fueling load growth, where the industrial-construction sector always plays a dominant role, accounting for approximately 68% of the district's total commercial electricity output. This characteristic makes Cu Chi a typical model to quantify the causal relationship between macroeconomic scale and energy demand.

B. Dataset Structure and Variable Definitions

To meet the training requirements of the forecasting algorithms, a continuous 16-year historical data series (2010 - 2025) was collected from the operational reports of Cu Chi Power Company and local statistical yearbooks. Four variables with absolute values were defined and incorporated into the simulation:

1. *Dependent variable (Y):* Total commercial electricity output (Dien_Tong) (unit: kWh).
2. *First exogenous variable (X1):* Average population size (*Dan_So*) (unit: People).
3. *Second exogenous variable (X2): Total net revenue of enterprises (Doanh_Thu_DN) (unit: Billion VND).*
4. *Third exogenous variable (X3): Total number of active enterprises (So_DN_Hoat_Dong) (unit: Enterprises).*

To provide a comprehensive view of the dataset's characteristics, the growth trends of commercial electricity output and revenue scale are visualized on the same reference frame.

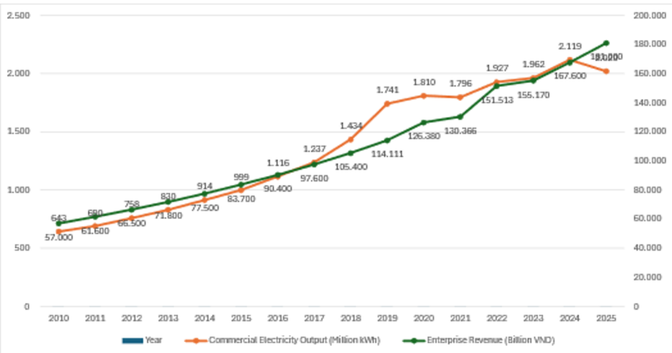

Figure 1. Comparison chart of absolute commercial electricity output and enterprise revenue in Cu Chi (2010-2025)

Observing (Figure 1), it can be seen that the increase in electricity consumption occurs covariantly and closely tracks the local economic expansion. Except for a slight decline in 2021 due to the objective impact of the pandemic, the data series quickly recovered its strong growth momentum in the following years.

C. Correlation Analysis and Data Pre-processing

Before establishing forecasting models, the input dataset is evaluated through a Pearson correlation matrix to screen and quantify the degree of linear association. The statistical test results are detailed in Table I. The Pearson correlation coefficient (*r*) is calculated as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Table I: Pearson correlation matrix among variables (2010 - 2025)

		Dien_Tong	Dan_So	Doanh_Thu_DN	So_DN_Hoat_Dong
Dien_Tong	Pearson Correlation	1	.986**	.968**	.977**
	Sig. (2-tailed)		.000	.000	.000
	N	16	16	16	16
Dan_So	Pearson Correlation	.986**	1	.987**	.991**
	Sig. (2-tailed)	.000		.000	.000
	N	16	16	16	16
Doanh_Thu_DN	Pearson Correlation	.968**	.987**	1	.989**
	Sig. (2-tailed)	.000	.000		.000
	N	16	16	16	16
So_DN_Hoat_Dong	Pearson Correlation	.977**	.991**	.989**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	16	16	16	16

Data from (Table I) indicates that all three exogenous variables exhibit extremely high positive correlation coefficients (approaching 1) with commercial electricity output. With a statistical significance level of P-value = 0.000, it can be affirmed that the impact of socioeconomic drivers on load demand has a solid scientific basis and does not arise from random noise. Thus, these three independent variables are retained to serve the model training process.

However, (Table I) also exposes a massive internal correlation among the independent variables themselves (e.g., the correlation coefficient between Population and Number of enterprises reaches 0.991). This uniform growth characteristic is the core cause of the multicollinearity risk. Therefore, a mandatory pre-processing step is to transform all absolute values of the 16-year series into natural logarithms (ln) format to narrow the variance, homogenize the scales, and control the influence of multicollinearity before feeding them into the forecasting algorithms.

IV. RESULTS AND DISCUSSION

A. Evaluation of Fitness and Error Levels on the Historical Dataset

The forecasting models including Multiple Linear Regression (OLS), Time Series (ARIMA), and Artificial Neural Network (ANN) were established and trained on the same SPSS software platform with a 16-year data series (2010 - 2025).

For the OLS method, the Analysis of Variance (ANOVA) yielded an F-value = 207.473 with a significance level of Sig. = 0.000, confirming that the model is statistically significant at a confidence level above 99%. The coefficient of determination $R2$ reached 0.981, indicating that 98.1% of the variation in commercial electricity output is reasonably explained by the increase in population size and enterprise development. Based on the Ordinary Least Squares algorithm, the actual load forecasting equation is formulated as follows:

$$lnE_t = -53{,}545 + 6{,}739 \times lnPop_t - 1{,}086 \times ln(Doanhthu_t) - 0{,}061 \times ln(SoDN_t)$$

Details on the contribution level and statistical significance of each independent variable in the equation are summarized in (Table II).

Table II. Estimation results of regression coefficients)

Variables	Coefficient B	Standard Error	t-value	P-value (Sig.)	VIF

 https://doi.org/10.1051/e3sconf/202672303003

Constant	-53.545	19.950	-2.684	0.020	-
Population (lnPop)	6.739	2.169	3.107	0.009	256.862
Revenue (lnDoanhThu)	-1.086	0.631	-1.720	0.111	195.198
Number of Enterprises (lnSoDN)	-0.061	0.668	-0.091	0.929	107.732

Data from (Table II) reveals a regular characteristic of macroeconomic data: the emergence of multicollinearity. All three independent variables have Variance Inflation Factors (VIF) exceeding the threshold of 100. The simultaneous growth among these variables disrupts the normal distribution of individual regression coefficients. Specifically, the Population variable holds a positive sign and possesses high statistical significance (Sig. = 0.009 < 0.05). **Statistically, this rejects the null hypothesis (H_0), confirming that population growth has a true and significant impact on electricity consumption.** However, the economic representative variables exhibit paradoxical negative signs (P-value > 0.05). However, based on the econometrics theory for time series, when the ultimate goal is total volume forecasting rather than micro-elasticity analysis, keeping these variables remains entirely valid. The overall interpolation power of the equation is not compromised.

To gain a quantitative perspective on the data-fitting capabilities of all three models, error metrics are summarized in (Table III).

Table III. Comparison of model errors on the training dataset (2010-2025)

Method	**Coefficient of Determination (R^2)**	**MAPE (%)**	**RMSE (Million kWh)**
Multiple Linear Regression (OLS)	0.981	4.12%	~ 106.8
Time Series (ARIMA)	0.968	4.16%	~ 92.1
Artificial Neural Network (ANN)	N/A	1.24%	N/A

Regarding the forecasting error calculation, due to the small sample size of the annual dataset (N=16), extracting a hold-out test set would severely degrade the regression accuracy. Thus, the software utilized all 16 observations to train the models. Data from Table III demonstrates that on this **in-sample training dataset**, the neural network (ANN) exhibits superior nonlinear recognition power with an **in-sample MAPE** extremely low at 1.24%. Meanwhile, the MLR/OLS and ARIMA models maintain in-sample error levels closely hovering around 4.1% along with acceptable absolute deviations (RMSE) for the district-level electrical scale. This error margin fully complies with stringent technical standards in power grid operational planning (MAPE < 5%). [12].

B. Future Scenario Analysis and Overfitting Risks

The fundamental difference between statistical methods and machine learning algorithms is starkly revealed during extrapolation forecasting for the 2026-2030 period. The input scenario is established based on the orientation that the local economy continues to sustain recovery momentum: enterprise revenue increases by an average of 8.5%/year, the number of enterprises grows by 5.5%/year, and the population rises by 2.5%/year. Based on this scenario, the OLS model and the ANN algorithm (optimized with a 1-hidden-layer architecture containing 2 neurons) produced two entirely contrasting forecasting trajectories in (Figure 2.).

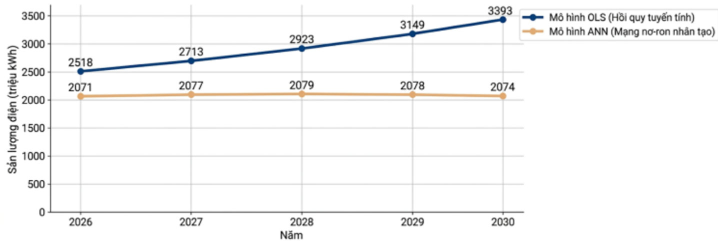

Figure 2. Comparison of the electricity output forecasting trajectory for the 2026-2030 period between the OLS model and the ANN

Relying on Figure 2, it is noticeable that for the OLS model, the regression equation generates a smooth growth trajectory, logically contiguous with the historical series. Commercial electricity output is forecasted to surge from 2,518 million kWh (in 2026) and hit the milestone of 3,393 million kWh by 2030. The average electricity growth rate achieves approximately 7.7% - 8.2%/year, fully consistent with the locality's industrial expansion orientation.

Conversely, the forecasting graph of the neural network (ANN) model portrays a severe logical failure. Instead of following the growth momentum of the input drivers, the ANN's forecasting trajectory flattens out. The predicted electricity output is sequentially saturated at the marks of 2,071; 2,077; 2,079; 2,078 and 2,074 million kWh for the 2026-2030 phase. This deviation exemplifies the state of Overfitting.

The collapse of the ANN model stems from the specific nature of the grid planning dataset. With annual statistical series, the training set encompasses only 16 observations (N = 16) [13]. This excessively small sample size prompts the algorithm to memorize local past noises rather than capturing the core long-term trends. Furthermore, the Hyperbolic tangent (Tanh) activation function at the neural network's hidden layer possesses saturation characteristics at both extremes [14]. Upon receiving future exogenous input variables (significantly larger than historical data), the output of the Tanh function asymptotes to a constant limit, causing the future forecasting scenario to saturate despite the continuously rising economic driving forces [15].

C. Discussion on Applicability and Explainability

Multi-model experimentation indicates immense technical risks if forcibly applying machine learning algorithms (ANN) to grassroots-level electrical data characterized by limited sample sizes. From another perspective, although the ARIMA algorithm detects the autoregressive structure ARIMA(0,1,0) helping bypass overfitting errors and maintain a good error margin (4.16%), it essentially remains a random walk model. It operates as a "black box" interpolating purely through mathematics, entirely ignoring the real macroeconomic impulses occurring in the locality.

In this context, the Multiple Linear Regression (OLS) method asserts itself as the most optimal and comprehensive tool. OLS not only safeguards the long-term growth forecasting trajectory, controlling the MAPE at 4.12%, but also yields the core value of Explainability. The quantitative outcome projecting load trends exceeding the 3 billion kWh threshold provides exceptionally robust scientific grounds, empowering grassroots technical managers to confidently defend investment portfolios, prioritizing capital allocation to upgrade 110kV substation capacities and renovate the 22kV distribution network to directly supply power to key industrial zones in the area.

V. CONCLUSION

This study has successfully addressed the medium-term electrical load forecasting problem at the local power management level by applying the Multiple Linear Regression (OLS) method. By directly quantifying the impacts of core socioeconomic drivers including population size, revenue, and the number of enterprises, the proposed model delivers a robust mathematical tool, replacing traditional qualitative forecasting approaches.

The most significant contribution of this paper is highlighting the technical limitations and practical risks of Artificial Intelligence algorithms (represented by ANN) when applied to the grid planning data structure. Comparative experimentation proves that, with the annual frequency of macroeconomic statistics leading to small sample sizes (N=16), machine learning algorithms fall into a severe overfitting state. The saturation of the activation function causes the ANN's future forecasting scenario to flatten out, entirely deviating from economic logic. In this context, the OLS model confirms its comprehensive superiority by sustaining low forecasting errors (MAPE achieving 4.12%), ensuring a reasonable interpolation trajectory corresponding to macroeconomic growth momentum, and providing absolute explainability in interpreting impact causes.

Regarding practical applications, the quantitative result forecasting the load scenario for Cu Chi district reaching 3,393 million kWh by 2030 closely aligns with the locality's socioeconomic development planning orientation and thoroughly satisfies the requirements in planning the operation of the distribution grid system as stipulated by the state energy management authorities (specifically Circular 19/2017/TT-BCT of the Ministry of Industry and Trade) [2]. This figure supplies solid scientific evidence for Cu Chi Power Company to proactively construct business plans, and petition competent authorities to prioritize investment capital allocation to upgrade 110kV substation capacities and expand the 22kV medium-voltage network to ensure a safe, continuous power supply to critical industrial parks. Furthermore, the simplicity of the OLS algorithm allows the Planning - Technical Department at the unit to effortlessly package the equation into standardized templates on common computational platforms to update, adjust weights, and run annual forecasts.

However, the study also notes certain limitations. Due to the covariant growth of macroeconomic variables, the model exhibits multicollinearity phenomena causing the individual regression coefficients to be distorted. Therefore, this equation is exclusively optimized for extrapolating the total future electrical load level, not intended for analyzing micro-policy elasticity (such as assessing the singular impact of a new enterprise on the grid).

To further enhance forecasting quality and broaden the foundation for deploying more complex machine learning algorithms, future studies should lean towards Data Disaggregation by monthly cycles to enlarge the observation sample size. Concurrently, incorporating additional variable clusters concerning environmental temperature and prevailing electricity prices will assist the model in comprehensively covering the agents influencing electrical consumption behaviors.

REFERENCES

[1] International Energy Agency - IEA (2020). *World Energy Outlook 2020*. IEA Publications, Paris.

[2] Ministry of Industry and Trade, Circular 19/2017/TT-BCT stipulating the content, method and sequence of implementing national power system load research (2017).

[3] N. Abu-Shikhah, F. Elkarmi, O. M. Aloquili, Medium-Term Electric Load Forecasting Using Multivariable Linear and Non-Linear Regression. Smart Grid and Renewable Energy, 2, 126-135 (2011).

[4] II. A. Samuel, C. F. Felly-Njoku, A. A. Adeyinka, A. A. Awelewa, Medium-Term Load Forecasting Of Covenant University Using The Regression Analysis Methods. Journal of Energy Technologies and Policy, 4(4), 4-16 (2014).

[5] E. C. Ashigwuike, A. R. A. Aluya, J. E. C. Emechebe, S. A. Benson, Medium term electrical load forecast of Abuja Municipal Area Council using Artificial Neural Network method. Nigerian Journal of Technology, 39(3), 860-870 (2020).

[6] L. Guo, L. Wang, H. Chen, Electrical Load Forecasting Based on LSTM Neural Networks. *International Conference on Big Data, Electronics and Communication Engineering*, 94, 107-111 (2019).

[7] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. MIT Press (2016).

[8] D. N. Gujarati, D. C. Porter, *Basic Econometrics* (5th ed.). McGraw-Hill Education (2009).

[9] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, *Multivariate Data Analysis* (7th ed.). Pearson Prentice Hall (2010).

[10] Huynh Tan Nguyen, Nguyen Van Luong, Application of ARIMA model in forecasting consumer price index CPI in Vietnam. Journal of Science, Hue University (2016).

[11] C. D. Lewis, *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth Scientific (1982).

[12] Le Huy Phuc et al., Application of statistical analysis to evaluate the reliability of input data sources to improve the quality of short-term electrical load forecasting on the HCMC power grid. Science and Technology Development Journal – Engineering and Technology, 2(4), 223-239 (2019).

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer (2006).

[14] Vu Huu Tiep, Basic Machine Learning. Science and Technics Publishing House (2018).

[15] Nguyen Thanh Thuy, Artificial Intelligence: Methods and Applications. Education Publishing House (2006).