International Conference on Actuarial Sciences (ICActS)
Available on-line at: **www.sicash.sliit.lk**

Forecasting Monthly Electricity Consumption for Energy Planning and Policy Development in Sri Lanka

Dineli Pathirana[1*], Chandima N. P. G. Arachchige[1]

[1]*Department of Statistics, University of Colombo, Colombo 03, Sri Lanka.*

Corresponding author*: dinelipathirana@gmail.com

Abstract

For effective energy planning, grid stability, and policy development especially in emerging nations like Sri Lanka accurate electricity consumption projections is essential. The goal of this project is to create a reliable model that can forecast the Ceylon Electricity Board's (CEB) monthly electricity consumption using a large dataset that includes macroeconomic variables, market indicators, peak demand, energy generation sources, and weather data. Autoregressive Distributed Lag (ARDL), Random Forest, and eXtreme Gradient Boosting (XGBoost) are the three models whose predicting performance is compared in this study. The most pertinent predictors were chosen using Recursive Feature Elimination with Cross-Validation (RFECV). Although XGBoost performed well throughout training, overfitting was a problem. ARDL was interpretable, however it was unable to detect long-term cointegration and could not represent non-linear connections. With the best accuracy and dependability on the test dataset without overfitting, Random Forest turned out to be the best model whereas Monthly Sales by Tariff in LKR, Fuel Cost by Power Stations in LKR, Electricity Generation from Thermal Coal in CEB (Gwh), Electricity Generation from Mannar Wind in CEB (Gwh), Day Peak Demand (MW), Night Peak Demand (MW), Average Monthly Rainfall (mm), and Gross Domestic Product (GDP) in LKR were the eight most important factors that were found to be involved in forecasting electricity consumption. On the test dataset, Random Forest, the best model chosen, had an accuracy of 77.34%, a Mean Absolute Percentage Error (MAPE) of 22.67%, a Root Mean Square Error (RMSE) of 12.62, and a Mean Absolute Error (MAE) of 11.11. However, the models might not be able to reflect long-term structural changes like the switch to electric vehicles or widespread adoption of renewable energy sources, and the study did not account for new elements like government policy reforms or energy efficiency initiatives. Nevertheless, the results show that machine learning, in particular Random Forest, can improve Sri Lankan electricity consumption predictions to aid in sustainable energy planning and policy choices.

Keywords: Time Series Modelling, RFECV, ARDL, Random Forest, XGBoost

Introduction

In developing nations like Sri Lanka, where grid dependability is essential and energy demand is changing quickly, accurate forecasting of electricity consumption is essential for efficient energy management, infrastructure development, and policy formation. This study aims to improve the Ceylon Electricity Board's ability to estimate electricity demand accurately, by analysing important variables like weather, economic trends, and changing energy production. The motivation of the study is to enhance Ceylon Electricity Board's (CEB) demand forecasting to facilitate the integration of renewable energy sources and tackle Sri Lanka's energy issues. The significance of the study rests in

https://doi.org/10.54389/GIOB1274

improving demand forecasting to enable data-driven energy decisions, optimize resources, and maintain economic stability. The Ceylon Electricity Board and Central Bank of Sri Lanka provided the monthly data used in this analysis, which is from January 2015 to December 2023. It supports precise demand forecasting and analysis of influencing factors by incorporating twelve time-based quantitative variables, including Electricity Consumption (Gwh) as the dependent variable and independent variables such as energy generating data (Hydro, Thermal Oil, Thermal Coal, and Mannar Wind in GWh), peak demand (Day and Night Peak Demand in MW), Revenue Billed by Tariff (in LKR million), Monthly Fuel Costs (in LKR million), Monthly Rainfall (in mm), Gross Domestic Product (GDP) in LKR. GDP in LKR variable was the only variable that was initially available in quarterly basis which was converted to monthly data using Cubic Spline Interpolation. Forecasting electricity demand is essential to attaining energy sustainability, especially in emerging nations like Sri Lanka. The literature examines a range of techniques, from contemporary Machine Learning (ML) algorithms to conventional econometric models, providing insightful information on estimating energy demand. Amarawickrama and Hunt (2007) applied Autoregressive Distributed Lag (ARDL) and other econometric techniques to analyse Sri Lanka's electricity demand, highlighting long-run income elasticity and forecasting future consumption patterns. Other researchers, such as Priyadarshana et al. (2021), have expanded ARDL to incorporate meteorological variables. However, these studies frequently leave out operational or tariff-related data in favour of concentrating mostly on causation or narrow variable scopes. Although machine learning techniques, particularly Random Forest and eXtreme Gradient Boosting (XGBoost), have demonstrated great short-term prediction abilities, there are still few uses for them in complicated, multidimensional dataset forecasting at the national level. Notably, "XGBoost effectively captures non-linear interactions while providing robustness to outliers and missing data" Sasikala et al. (2024). Inconsistencies in the model's generalizability and interpretability across settings continue, despite the growing usage of weather and economic variables. Domain-specific operational data, such as peak demand or fuel-type energy generation, are frequently not integrated into the examined studies. This disparity emphasizes the necessity of comprehensive models that incorporate technological, environmental, and economic aspects. To create reliable, policy-relevant forecasting frameworks that are suited to the real-time utility requirements of developing nations, more research is necessary. This study's primary objective is to create a forecasting model that can accurately estimate the CEB's monthly electricity consumption. Additionally, it seeks to pinpoint the main factors influencing CEB's electricity usage as the secondary objective. The study also contrasts the efficacy of conventional statistical techniques like ARDL with machine learning techniques like Random Forest and XGBoost as another secondary objective. Prior research frequently ignores operational or tariff-related information in favor of concentrating only on causality or a small number of variables. Despite their outstanding short-term prediction capabilities, machine learning methods like as Random Forest and XGBoost are not widely used for forecasting complicated, multidimensional national datasets. By specifically asking the following research questions, this study fills in these gaps: Is it possible for a model to predict CEB's monthly power use with any degree of accuracy? Which elements have the biggest effects on usage? Do machine learning approaches perform better than traditional methods such as ARDL? According to the associated hypothesis, machine learning techniques perform better than conventional methods, operational and economic aspects are important, and integrated models increase forecast accuracy.

Materials and Methods

The research technique, implemented using RStudio including data collecting, preprocessing, interpolation, and analytic approaches, is described in this part. With the help of pertinent terminology and equations, it describes in detail how time series and machine learning models are used. Preliminary Analysis started off with time series plots to identify any trends, seasonality, cyclic patterns and irregular fluctuations considering the original dataset with 108 observations spanning from January 2015 to December 2023 on all 11 variables time-based quantitative collected from the Central Bank of

Sri Lanka (CBSL) and CEB. Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots were drawn initially on all variables. ACF and PACF plots to determine lag lengths for both dependent and independent variables mentioned by Montgomery et al. (2015). Moving forward, seasonality was tested using seasonal unit root test and removed with the help of Seasonal-Trend decomposition utilizing Loess (STL). "LOESS is Local regression" Robert et al. (1990), contributing to a decomposition method robust to anomalies. The dataset was split to an 80-20 split was used to separate the data into a training set, which included 86 monthly observations from January 2015 to February 2022, and a testing set, which included 22 observations from March 2022 to December 2023. Furthermore, stationarity was tested on all variables on the training set using ADF tests and cross correlation matrix was plotted afterwards to identify similarities between time series variables at different lags.

Recursive Feature Elimination and Cross-Validation Selection (RFECV) was used to determine the most significant features from the 10 independent variables using the training set. Afterwards, integration orders were obtained and due to the difference in integration orders on variables ARDL bound test for cointegration was tested on the selected significant variables from RFECV. Since the above test resulted in no cointegration between the variables in this instance, an error correction model (ECM), which is commonly employed to capture long-term relationships, is not necessary. To adequately capture the immediate, short-term dynamics among the significant variables, a short-run ARDL model with first-differenced variables is used. The Akaike Information Criterion (AIC), which penalizes the inclusion of superfluous lags or variables to balance model fit and parsimony, served as the basis for the model selection process. To clearly distinguish between short-term impacts and long-term equilibrium correction, the standard ARDL model can be reparametrized into an Unrestricted Error Correction Model (UECM) as quoted by Pesaran, M. H., Shin, Y., & Smith, R. J. (2001).

$$\Delta y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \sum_{j=0}^{q} \delta_j \Delta x_{t-j} + \lambda EC_{t-1} + \epsilon_t \text{ ---------- (1)}$$

In the ARDL model framework, Δ denotes the first difference capturing short-run changes, where y_t is the dependent variable and x_t represents an independent variable, the coefficients β_i and δ_j reflect short-run dynamics, λ is the error correction coefficient indicating the speed of adjustment back to equilibrium, $EC_{t-1} = y_{t-1} - \theta x_{t-1}$ is the error correction term (ECT) representing the lagged long-run equilibrium relationship, and ε_t denotes the stochastic error term. Model diagnostics and stability tests applied to the ARDL model, including tests for autocorrelation, heteroscedasticity, normality, model specification, structural breaks and predictive causality between variables such as Breusch-Godfrey Test, Breusch-Pagan Test, Ljung-Box Test, Ramsey RESET Test, Shapiro-Wilk Test, Cumulative Sum Control Chart (CUSUM) Test, CUSUM of Squares Test, Moving Sum Test (MOSUM) Test and Granger Causality Test. Furthermore, a potent supervised machine learning approach was used called Random Forest on the significant variables which build a Forest by combining several decision trees. Regression and classification are two uses for it. In essence, a random forest classifier or regressor is a bagging method. To have better performance and to make the model faster, hyperparameters such as mtry, bootstrap, maxdepth, nodesize and ntree are used in random forests. Finally, as the last approach the significant variables were fitted using gradient descent to minimize prediction errors, the scalable and potent machine learning technique known as XGBoost which creates models one after the other. It corrects the residuals of the ensemble's earlier forecasts by adding a new tree at each iteration. Before projections are derived, fitted models should be assessed using a forecasting performance metric. Since erroneous models may generate issues with fitted models, forecasting accuracy is crucial in time series analysis. To ensure a thorough evaluation of both model fit and predictive capability, the performance of forecasting models was assessed using a variety of indicators, including R2 (coefficient of determination), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Accuracy.

Results

From January 2015 to December 2023, the time series plots showed clear patterns, seasonality, and erratic fluctuations for every variable. Peak demand, GDP, and electricity consumption all exhibit robust increasing patterns, with sporadic hiccups brought on by the COVID-19 pandemic. Rainfall, coal, hydro, and other variables show strong seasonality, whereas wind and oil generation show unpredictable trends and operational shocks. Despite seasonal decomposition, autocorrelation is nevertheless evident in many series, as seen by ACF and PACF plots, which show continuous non-stationarity. To ensure model correctness and robustness in capturing time-dependent dynamics, these visual patterns highlight the significance of checking for stationarity and performing the proper transformations prior to forecasting. Seasonal Unit Root Test identifying important seasonal components in variables like GDP, coal and hydropower generation. ADF tests are used to evaluate stationarity on the training dataset after removing seasonality, and the results indicate that just five variables including GDP, rainfall, hydro, thermal oil and coal generation are stationary, with the remaining variables are not stationary. Most time series, especially those pertaining to energy consumption and peak demand factors, show significant persistence and non-stationarity, according to ACF and PACF plots. In addition to, cross-correlation matrix illustrates the substantial positive correlations between GDP and electricity consumption.

The study's main goal was to predict Sri Lanka's electricity consumption using both sophisticated machine learning models and conventional econometric methods. RFECV was initially used for feature selection with 10-fold cross validation, and eight important variables that maximized prediction accuracy while reducing complexity were found. Metrics including monthly sales by tariff, night and day peak demand, GDP, fuel cost by power stations, and the many types of electricity generation (mannar wind, thermal coal, and hydro) were among them. In line with Hammad et al. (2024), who also showed the efficacy of RFE in conjunction with Random Forests for energy forecasting, this selection technique produced a well-balanced model performance with an R^2 of 0.7996 and RMSE of 10.50. ADF tests were used to determine integration orders and evaluate stationarity. For ARDL modelling, most variables were either I (0) or I (1). In accordance with the integration order analysis approach, mannar wind generation, which was determined to be I (2) and borderline stationary at 10%, was carefully incorporated into the model. The ARDL Bound Test showed a high p-value of 0.9861 and an F-statistic of 0.81188, indicating no long-term cointegration. 95.08% of the variation in electricity use was described by the chosen ARDL model, which was selected using the auto_ardl() function and AIC (R2 = 0.9508). Several diagnostic tests validated the model's dependability and confirmed its resilience. The Breusch-Pagan Test (p = 0.821) verified homoscedasticity, and the Breusch-Godfrey Test (p = 0.855) revealed no serial correlation. Ramsey's RESET Test (p = 0.8562) and the Ljung-Box Test (p = 0.879) revealed no problems with functional form or residual autocorrelation. The Shapiro-Wilk Test, however, showed non-normal residuals (p = 0.0006). While other factors such as hydro, coal, day and night peak demand, fuel cost, tariff, and GDP did not exhibit a causative link (p > 0.05), Granger causality analysis revealed wind generation to be a significant predictor of electricity consumption (F = 2.7817, p = 0.046). For machine learning-based prediction, Random Forest and XGBoost were used in addition to the ARDL model. Grid search and 5-fold cross-validation were used to pick 500 trees with optimum mtry = 2 for training the Random Forest model. This arrangement produced the best results, with an R2 of 0.7947 and an RMSE of 11.04. Among the most important predictors found were hydro and thermal coal generation and day peak demand. Strong non-linear effects of these variables on power consumption were shown through partial dependence plots. Concurrently, 450 boosting rounds, a maximum depth of 9, and a learning rate (eta) of 0.1 were used to fine-tune the XGBoost model. Other values were also set, such as subsample = 0.7, min_child_weight = 1, and colsample_bytree = 0.5. With a training RMSE of 0.0005451, which is close to zero, the XGBoost model demonstrated exceptionally high predicted accuracy on the training set.

Model Comparison

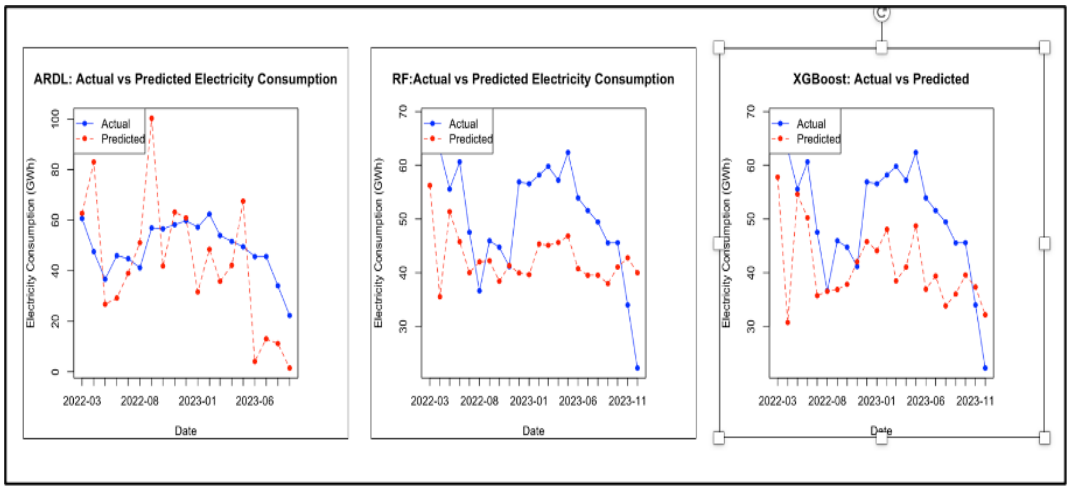

Figure 1: *Comparison of Actual vs Predicted Electricity Consumption*

Figure 1 demonstrates how the mid-2022 and 2023 ARDL projections are significantly different from the actual consumption, although the forecasts from RF and XGBoost are more precise.

Table 1: *Summary of Performance Measures on Fitted Models*

Model	Dataset	MAPE	MAE	RMSE	Accuracy
ARDL	Test	40.55%	18.27	22.05	59.45%
ARDL	Train	23.09%	3.89	5.22	76.91%
RF	Test	22.67%	11.11	12.62	77.34%
RF	Train	21.64%	3.21	4.87	78.36%
XGboost	Test	21.3%	10.99	13.04	78.7%
XGboost	Train	0.003%	0.0004	0.0005	99.99%

Table 1 shows the performance metrics of the fitted models on both training and test datasets. ARDL shows relatively poor predictive performance, with a high test MAPE of 40.55% and a test accuracy of 59.45%, indicating limited ability to capture the underlying time series patterns. XGBoost and Random Forest (RF) both perform better than ARDL, with XGBoost having the lowest test MAPE (21.3%) and the highest test accuracy (78.7%), but XGBoost's near-perfect training performance suggests possible overfitting, underscoring the need for appropriate diagnostic testing to balance fit and predictive accuracy.

Discussion

The results show that machine learning models, especially Random Forest, outperformed ARDL in forecasting Ceylon Electricity Board's consumption by capturing complex non-linear relationships. RFECV improved model accuracy by selecting key predictors like monthly sales by tariff, night and day peak demand, GDP, fuel cost by power stations, electricity generation by mannar wind, thermal coal, and hydro. XGBoost overfitted despite good training accuracy. Limitations include absence of real-time data and uncertainty quantification. Future research should explore adaptive learning and probabilistic forecasting to enhance responsiveness and decision-making under uncertainty.

Conclusions

When predicting electricity consumption using 8 important indicators like monthly sales by tariff, night and day peak demand, GDP, fuel cost by power stations, electricity generation by mannar wind, thermal coal, and hydro, machine learning models in particular, Random Forest performed better than XGBoost and conventional model like ARDL. This demonstrates how machine learning is better at improving Sri Lanka's sustainable energy planning. This demonstrates how machine learning could improve Sri Lanka's plans for renewable energy. These results suggest that to improve generation and distribution, authorities should think about incorporating machine learning-based forecasting into national energy planning. To guarantee an effective, dependable, and sustainable supply of electricity, these predictive insights can also be used to guide policies like dynamic tariff adjustments, enhanced fuel cost monitoring, and investments in renewable energy infrastructure.

References

Amarawickrama, H. A., & Hunt, L. C. (2008). *Electricity demand for Sri Lanka: A time series analysis*. *Energy, 33*(5), 724–739.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*(5), 1287–1294.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica, 46*(6), 1303–1310.

Hammad, M., Khan, A., & Zhang, Y. (2024). Enhancing energy load and price forecasting using Random Forests and recursive feature elimination. Energy Informatics Journal, 18(2), 135–148.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting* (2nd ed.). Wiley.

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics, 16*(3), 289–326.

Priyadarshana, K., Weerathunga, P., & Jayasundara, J. M. (2021). Climate change and electricity demand in Sri Lanka: An extended ARDL approach. *Energy & Environment, 32*(6), 1184–1202.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological), 31*(2), 350–371.

Robert, C., William, S., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–73.

Sasikala, P., Manikandan, T., & Vasanthi, V. (2024). Forecasting electricity demand using XGBoost: A robust approach for outlier resilience. *Journal of Cleaner Energy Technologies, 12*(1), 45–52.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591–611.