INTRODUCTION

Time series forecasting is a critical component of many economic, financial, and scientific applications, where accurate predictions can lead to significant improvements in decision making and operational efficiency. Traditional models, such as the Auto-Regressive Integrated Moving Average (ARIMA), have been widely used due to their robustness in capturing linear relationships and trends in time series data. However, these models often fall short when dealing with non-linear patterns that are increasingly present in complex systems like financial markets, weather forecasting, and energy demand management.

To address these limitations, researchers have turned to machine learning techniques, such as Multi-Layer Perceptrons (MLP), which can model complex non-linear relationships hidden in the data. Despite their flexibility, MLPs alone can sometimes overlook the seasonal and trend components that ARIMA models handle well. This has led to the burgeoning interest in hybrid models that combine the strengths of statistical and machine learning approaches to enhance forecasting accuracy and reliability.

Hybrid models, specifically those integrating SARIMA (Seasonal ARIMA) and MLP, offer a promising avenue for capturing both linear and non-linear dynamics in time series data. The SARIMA model extends ARIMA by accounting for seasonality, an essential feature for many practical applications such as electricity load forecasting and inventory management, where patterns repeat over a known cycle. By coupling SARIMA with MLP, the hybrid approach aims to leverage SARIMA's effectiveness in handling seasonal variations and the powerful pattern recognition capabilities of MLP to improve overall forecast performance.

Recent studies have demonstrated the superiority of hybrid models over their standalone counterparts. For instance, the weighted MLP-ARIMA model, as discussed in recent literature, outperforms traditional ARIMA and MLP models by integrating the predictive capabilities of both, ensuring more accurate and reliable forecasts (Hajirahimi and Khashei, 2020)​(JIEMS\_Volume 7\_Issue 2\_…)​. Moreover, the concept of adjusting the weight of predictions from each model component in the hybrid framework allows for fine-tuning the model's response to changes in data dynamics, further enhancing the model's adaptability and accuracy.

This paper builds upon this foundation by exploring the specific hybrid combinations of SARIMA-MLP and MLP-SARIMA. Each configuration is designed to test how different sequences of these models affect the forecasting accuracy, providing insights into the optimal structuring of hybrid models for various time series forecasting challenges. The goal is to ascertain a robust framework that can be effectively applied across different domains, offering a substantial improvement over existing methodologies.

THEORICAL FOUNDATIONS

Trend Analysis

To assess the presence of trend and seasonal components in time series data, the Cox-Stuart and Kruskal-Wallis tests are utilized, respectively (Santos et al., 2021). If either component is detected, a common approach is to apply successive differencing to the original series to achieve stationarity. Removing the trend component typically involves taking the first difference of the series. Generally, two differences are sufficient. Deterministic seasonality can be addressed by differencing the series using the moving average method, tailored to the seasonal period 's' observed in the data.

Additive and Multiplicative Models for Time Series Analysis

The next step was to determine the appropriate model equation to relate the components to the variable, specifically choosing between an aditive  or multiplicative  model. In these models, ​ represents the value of the time series, ​ is the trend component, is the seasonal component, and ​ is the random component, all at time . In an additive model, the amplitude of the seasonal pattern remains constant over time, regardless of changes in the trend.

Data Transformation Necessity

The presence of a linear relationship between amplitude and mean values in a time series, as noted by Ferreira et al. (2020), indicates heteroscedasticity, necessitating data transformation to stabilize variance. The Box-Cox transformation is particularly effective in such scenarios as it normalizes the distribution and stabilizes variance across the data. By transforming the data to meet the constant variance assumption required for many statistical analyses, this method enhances the reliability and accuracy of subsequent analyses and modeling, ensuring that statistical inferences and predictions are valid and robust [BOX, JENKINS 1976].

SARIMA MODEL

Recent studies have utilized statistical techniques to analyze historical time series data on wind energy across various regions, highlighting significant contributions. Alencar (2018) developed a forecasting model for wind speed in wind energy generation using a hybrid methodology that incorporates the SARIMA model and artificial neural networks. Fernandes (2018) performed short-term wind power forecasting based on spectral analysis and time series decomposition. Silva (2018) examined and compared three methods for reducing the dimensionality of time series to analyze extreme wind values. Silva (2017) conducted short-term wind forecasting using statistical modeling in areas with wind energy potential in Northeast Brazil. Malta (2009) adjusted a time series model to predict wind speeds to assess the feasibility of wind farm installations.

Building on these methodologies, the ARIMA and SARIMA models, known respectively as the Box-Jenkins method, serve as foundational techniques for short-term forecasting in time series analysis. The ARIMA model is denoted as ARIMA(p, d, q), where 'p' represents the autoregression component indicating the dependency of the series on its previous values, 'd' denotes the integration part used for non-stationary series, and 'q' represents the moving average component reflecting the dependency on past error terms. The SARIMA model extends ARIMA by incorporating seasonality, making it exceptionally suitable for datasets like wind speeds that exhibit seasonal patterns. The process involves four key steps: model identification, parameter estimation, diagnostic checking, and forecasting. Initially, the series undergoes tests for stationarity and seasonality, followed by model identification through correlogram analysis. After estimating the parameters, the models are evaluated through diagnostic checks to ensure adequacy, and if satisfactory, they are used for forecasting. This comprehensive approach allows for a robust analysis of time series data, enhancing the predictive accuracy and reliability of models used in wind energy forecasting (Syahrini, 2023.

MLP MODEL

The Multilayer Perceptron (MLP) model, a machine learning framework, is extensively recognized for its versatility and accuracy in forecasting time series data. This model excels in approximating any continuous, nonlinear, differentiable, and bounded function. Structurally, the MLP comprises an input layer, one or several hidden layers, and an output layer. It employs artificial neurons to transmit data sequentially from one layer to the next, with hidden layers processing input data through nonlinear functions specific to the study focus (Qureshi, Daniyal, & Tawiah, 2022). The MLP model operates on neural network principles using adaptive weights to manage input data. The neuron's output, denoted by YYY, results from the input xxx, weight www, and bias bbb, combined using an activation function, shown as  (Wicaksana et al., 2022). Activation functions such as the sigmoid , hyperbolic tangent , and rectified linear unit  are integral to the model. Setting up an MLP involves data selection, training, and model validation.

SARIMA-MLP Model

This section discusses the hybrid forecasting model that combines the Seazonal Autoregressive Integrated Moving Average (SARIMA) with the Multilayer Perceptron (MLP) specifically for solar radiation forecasting. The SARIMA model, primarily linear, is traditionally employed to predict time-series data and tends to produce nonlinear residuals. To address these nonlinear residuals, the MLP, a nonlinear model, is utilized, which results in residuals that typically exhibit a linear structure (Rajalakshmi & Vaidyanathan, 2022). Initially, the SARIMA model processes the linear aspects of solar radiation data. Subsequently, the MLP trains on the nonlinear residuals produced by the SARIMA model, effectively managing the nonlinear error components. The SARIMA model's initial output, denoted by ​ at time , and the residual ​ retain nonlinear patterns that are then modeled using the MLP in the subsequent stage. This MLP model for the residuals, represented by equation

= , calculates the predicted value at time t from the MLP model on the residual data, with  being the random error and f the nonlinear function defined by the MLP.

MLP-SARIMA Model

The methodology detailed in the document centers on a sophisticated hybrid forecasting model known as the weighted MLP-ARIMA series hybrid model. This model, developed by Zahra Hajirahimi and Mehdi Khashei (2020), enhances traditional forecasting techniques by strategically decomposing time series data into its fundamental components—nonlinear and linear—and optimally weighting the forecasts derived from each. Here's a concise overview of the process:

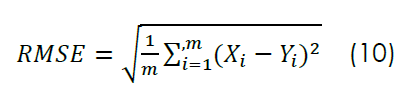
1. **Decomposing Time Series**: The initial step involves splitting the time series into its nonlinear and linear components, treating the actual data yty\_tyt​ as a sum of these two distinct parts, yt=Nt+Lty\_t = N\_t + L\_tyt​=Nt​+Lt​.
2. **Component-Specific Modeling**: The MLP model is first employed to handle all nonlinear patterns present in the data. Subsequently, the ARIMA model captures the remaining linear relationships by utilizing the residuals from the MLP, ensuring that all aspects of the data are thoroughly analyzed.
3. **Optimal Weight Assignment**: This model distinguishes itself by using the OLS algorithm to determine the exact weights for the forecasts generated by both the MLP and ARIMA models, thus maximizing the efficacy of each model based on its forecasting accuracy.
4. **Constructing the Final Hybrid Forecast**: The final forecast is a composite of the individually weighted predictions from the MLP and ARIMA models, producing a nuanced forecast that leverages both models' strengths.

This methodological approach proposed by Hajirahimi and Khashei offers a refined technique for time series forecasting, allowing for an enhanced predictive performance by effectively balancing the contributions of nonlinear and linear predictors.

MODEL EVALUATION

The coefficient of determination, R², quantifies the percentage of variance in the dependent variable that is predictable from the independent variables (Chicco, Warrens, & Jurman, 2021). When identifying outliers, the Mean Squared Error (MSE) is advantageous due to its ability to place greater emphasis on such data points, as it assigns higher weights to these outliers. Additionally, MSE is closely linked to the Root Mean Squared Error (RMSE) through a monotonic relationship involving the square root, meaning that ranking models based on MSE is equivalent to ranking them based on RMSE. This study utilizes both R² and RMSE to assess the model's performance. The R² is calculated as shown in equation 9, where YiY\_iYi​ represents observed values, XiX\_iXi​ denotes predicted values, and Y‾\overline{Y}Y is the mean of all observed data. Similarly, RMSE, which measures the square root of the average of the squared differences between predictions and actual observations, is computed as per equation 10, and is used to gauge model accuracy, with lower RMSE values indicating higher model precision. This measure was specifically applied to determine the best model for estimating solar radiation, considering the number of periods used in the calculations.





METHOD FRAMEWORK

Dataset and Time Series

The time series discussed in this article comprises historical data on electric energy production from wind sources, collected monthly by the National Electric System Operator (ONS, 2024). This series encompasses 210 monthly observations measured in megawatt-hours (MWh) spanning from January 2007 to June 2024. The data was segmented into two subsets; the first, from January 2007 to December 2022, was used for model calibration, and the second, from January 2023 to June 2024, was used for forecast evaluation. The forecasting horizon extended from July 2024 to December 2025, discribed on Figure 1. A time series is defined as a sequence of observations spaced at regular time intervals, representing a stochastic process thought of as a family of random variables defined in a common probability space. This study is exploratory and quantitative, employing statistical techniques for data analysis and modeling as described by Pereira, Shitsuka, Parreira, and Shitsuka (2018).

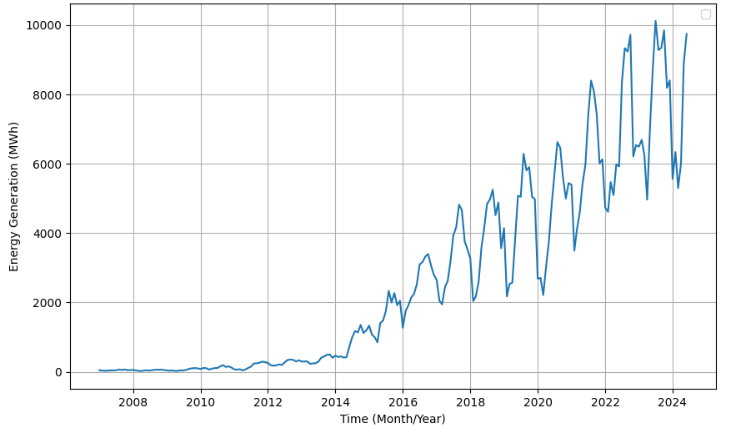


Figure 1: Monthly series of electrical energy production generated by the wind array in the period from Jan/2007 to Jun/2024.

Model Verification: Additive vs. Multiplicative

To determine whether the model is additive or multiplicative, an amplitude versus mean plot was created (Figure 2). Analysis of this plot revealed a positive relationship between the means of the observations of subgroups of the original series and the amplitude of these subgroups, with a slope coefficient of 0.7507, estimated through simple linear regression. The obtained p-value was less than 0.01, rejecting the null hypothesis that the slope coefficient is zero at the 5% significance level. Thus, it is concluded that the model is multiplicative.

REFERENCES

Alencar, D. B. (2018). Modelo híbrido baseado em séries temporais e redes neurais para prevenção da geração de energia eólica. Tese de doutorado,

Universidade Federal do Pará, Belém, PA, Brasil.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Fernandes, L. C. (2018). Previsão de potência eólica de curtíssimo prazo baseada na análise espectral e decomposição da série temporal. Dissertação de

mestrado, Universidade Federal de Pernambuco, Recife, PE, Brasil.

Ferreira, R. A., Miranda, V. F. L., Santos, P. M., & Sáfadi, T. (2020). Um estudo sobre a evolução de óbitos por câncer de mama no Brasil usando modelos de

séries temporais. *Research, Society and Development*, 9(12), 1-20.

Z. Hajirahimi and M. Khashei, "Weighted MLP-ARIMA series hybrid model for time series forecasting," J. Ind. Eng. Manag. Stud., vol. 7, no. 2, pp. 187-201, 2020. DOI: 10.22116/JIEMS.2020.205148.1307

Malta, C. S. (2009). Estudos de séries temporais de vento utilizando análises estatísticas e agrupamento de dados. Trabalho de conclusão de curso,

Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.

M. Qureshi, M. Daniyal, and K. Tawiah, “Comparative Evaluation of the Multilayer Perceptron Approach with Conventional ARIMA in Modeling and Prediction of COVID-19 Daily Death Cases,” J. Healthc. Eng., vol. 2022, 2022, doi: 10.1155/2022/4864920.

V. Rajalakshmi and S. G. Vaidyanathan, “Hybrid Time-Series Forecasting Models for Traffic Flow Prediction,” Promet - Traffic - Traffico, vol. 34, no. 4, pp. 537–549, 2022, doi: 10.7307/ptt.v34i4.3998.

Santos, P. H. A. B., Delfino, O. A. S., Santos, R. V. R., & Nascimento, M. (2021). Ajuste de um modelo de séries temporais para prever a precipitação

pluviométrica. *Research, Society and Development,* 10(6), 1-11.

Silva, H. B. (2018). Técnicas para redução de dimensionalidade de séries temporais e detecção de velocidades extremas do vento para geração de energia

eólica. Dissertação de mestrado, Universidade Federal de Pernambuco, Recife, PE, Brasil.

I. Syahrini, R. Radhiah, W. F. Damanik, N. Sciences, U. S. Kuala, and B. Aceh, “Application of the Autoregressive Integrated Moving Average (ARIMA) Box-Jenkins Method in Forecasting Inflation Rate in Aceh,” *Transcendent J. Math. Appl. ISSN*, vol. 2, no. 1, pp. 27–33, 2023.

H. S. Wicaksana et al., “Air Temperature Sensor Estimation on Automatic Weather Station Using ARIMA and MLP,” vol. 14, no. 2, 2022.