International Journal of Mathematics and
Computer Science, **19**(2024), no. 1, 195–203

A Comparative Study of Traditional Methods and Hybridization for Predicting Non-Stationary Sunspot Time Series

Muzahem M. Al-Hashimi[1], Heyam A. A. Hayawi[2], Mowafaq Al-Kassab[3]

[1,2] Department of Statistics and Informatics
College of Computer Science and Mathematics
University of Mosul
Mosul, Iraq

[3]Department of Mathematics Education
College of Education
Tishk International University
Erbil, Iraq

email: muzahim_alhashime@uomosul.edu.iq

(Received July 15, 2023, Accepted August 16, 2023,
Published August 31, 2023)

Abstract

Predicting sunspot numbers presents ongoing challenges in forecasting, including non-stationary patterns and unclear fluctuation dynamics. This study compares traditional methods and hybrid models, incorporating machine learning techniques, to predict monthly mean sunspot numbers (MMSNs) from January 1, 1900, to December 31, 2022. Among the traditional methods, ARIMA(5,0,4) demonstrated performance with an MSE of 580.949, RMSE of 24.103, MAE of 17.19, and MAPE of 0.511. However, the proposed hybrid model, which combines ARIMA(5,0,4) with additive regression (AR) using Regression by Discretization (RegbyDisc) based on J48, achieved markedly superior forecasting accuracy with an MSE of 114.653, RMSE of 10.708,

Key words and phrases: Sunspots, ARIMA, Additive Regression, Regression by Discretization, J48, Hybrid Model.
AMS (MOS) Subject Classifications: 60, 62.
ISSN 1814-0432, 2024, http://ijmcs.future-in-tech.net

MAE of 6.441, and MAPE of 0.438. We employed this hybrid model to forecast sunspot numbers from January 2023 to December 2025.

1 Introduction

Real-world systems can be complex and difficult to understand, making measurements a crucial component of studying such systems. Time series data is a commonly used measurement method in fields such as weather and environment where accurate forecasting is essential [1]. Sunspots are the primary metric used to measure solar activity. They are characterized by dark areas on the solar disk [2] and have two primary advantages: they are easily visible indicators of solar activity and there is a publicly available historical record of sunspot numbers dating back to 1749 [3], making them a valuable resource for studying the effects of solar activity on Earth's environment.

Scientists have been studying historical data on sunspots and associated solar activity for over a century, uncovering a wealth of information about the sunspot cycle [4]. However, predicting the time series of sunspot numbers is a challenging area of forecasting that remains an open challenge. Despite its importance, this time series poses several practical difficulties that must be addressed, including its nonstationary nature and the unclear dynamics underlying the fluctuations in cycle amplitude. It is not clear whether the data represents a noisy limit cycle or whether complex dynamics are at play [5]. The sunspot cycle profiles exhibit strong deviations from a sinusoidal shape, with a peak that is pronounced and asymmetrical. Additionally, the statistical distribution of sunspot numbers departs significantly from a Gaussian distribution [6].

Hybrid Machine Learning (HML) is a two-stage process that integrates multiple techniques to address model selection challenges and improve time series forecasting precision [7]. HML recognizes that real-world time series often exhibit linear and nonlinear patterns, making it difficult to choose the most appropriate technique for a particular situation[7, 8].

This paper proposes a novel hybrid model that combines ARIMA and AR method, incorporating RegbyDisc based on J48, for sunspot number forecasting.

2 Sunspots Dataset

In this study, MMSN data were collected from the World Data Center Sunspot Index and Long-term Solar Observations (WDC-SILSO). The data spans January 1900 to December 2022. The time series covers a period of 123 years, comprising 1476 months, and displays cyclic patterns where the observed values fluctuate in regular periods. Figure 1 displays the time series plot of MMSNs from January 1900 to December 2022. To forecast the MMSN, the dataset was split into two subsets: a training set comprising the initial 90% of rows and a test set encompassing the remaining 10% of rows.

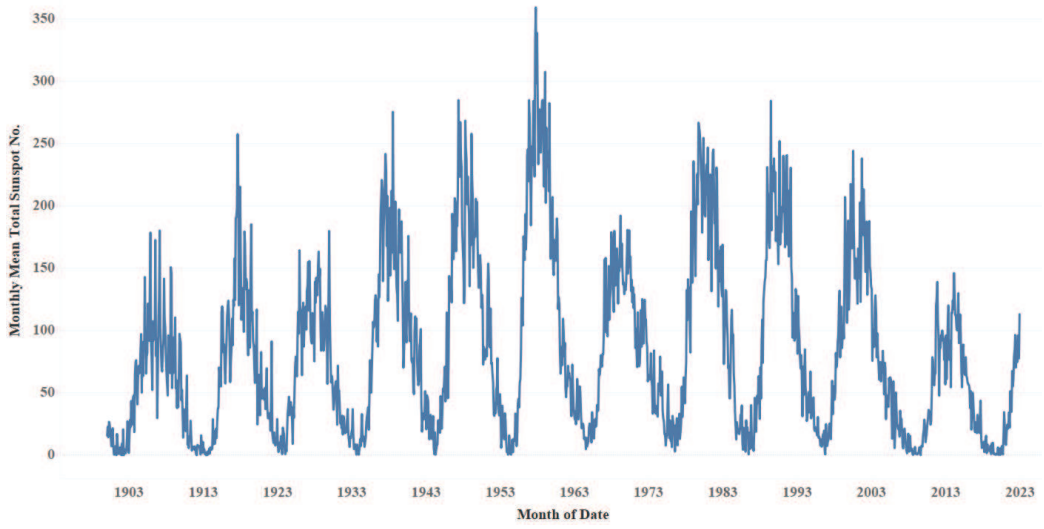

Figure 1: The MMSN spanning from January 1, 1900, to December 31, 2022

3 Methods

3.1 Traditional Methods

Time series analysis involves a variety of statistical techniques that have been utilized for several decades, known as traditional methods. ARIMA, in particular, is a widely used and important method for forecasting in diverse fields. Notably, many traditional methods can be expressed in terms of ARIMA with varying orders and coefficients, making ARIMA a unifying framework that encompasses several traditional techniques.

3.1.1 ARIMA (p, d, q) model

An ARIMA (p, d, q) model can be defined as a regressive Integrated Moving Average Autoregressive model. It is distinguished by three parameter types [9]. The 'p' parameter represents the number of autoregressive terms (AR), which predict the variable based on its own lagged values. The 'd' parameter signifies the number of differences (order) introduced to eliminate non-stationarity. The 'q' parameter indicates the number of moving average terms (MA), which predict the variable based on previous regression errors.

3.2 Machine Learning Methods

3.2.1 Additive regression (AR)

Additive regression as a nonparametric regression technique [10], enhances the performance of weak prediction models based on a specified criterion. It achieves this by aggregating contributions obtained from other models. Instead of building the base models independently, most learning algorithms for AR focus on ensuring their mutual complementarity and strive to form an ensemble of base models that collectively improve predictive accuracy [11].

3.2.2 Regression by Discretization (RegbyDisc)

It refers to a regression approach that involves discretizing the class attribute into a pre-defined number of bins using equal-width discretization. This methodology can utilize a distribution classifier or any classifier on a replicated or modified dataset [11].

3.2.3 J48

J48 is a Java-based implementation of the C4.5 decision tree algorithm primarily used for classification tasks. It assigns class labels to instances in supervised datasets based on their attribute values. J48 constructs decision trees by analyzing attribute values in the training set and selecting the attribute that best categorizes instances [12].

3.3 Hybrid Model

The hybrid model Y_t consists of two components: the linear component L_t and the nonlinear component N_t, i.e., $Y_t = L_t + N_t$. The linear component L_t is acquired through an ARIMA model. To calculate the model's residual at time t, we use the formula $\varepsilon_t = Y_t - \hat{L}_t$, where $\hat{L}_t$ is the predicted value of the ARIMA model at time t. The residuals obtained are then modeled using AR with RegbyDisc based on J48. They can be expressed as $\varepsilon_t = f(\varepsilon_{t-1}, \ldots, \varepsilon_{t-n}) + \Delta_t$ [13]. The AR with RegbyDisc based on J48 captures the nonlinear behavior of the data through a function denoted as f. The model also includes a random error term Δ_t. As a result, the combined forecast can be expressed as $\hat{Y}_t = \hat{L}_t + \hat{N}_t$, where $\hat{N}_t$ represents the predicted value of ε_t. [13, 14].

4 Results and Discussion

The conducted study focused on developing a reliable hybrid prediction model for sunspot time series data. The initial step involved selecting an appropriate traditional model, which led to a thorough analysis of forty-two forecasting methods. These methods encompassed a range of approaches such as Linear Regression with Autoregressive Errors, Seasonal Exponential Smoothing, Winter Method-Additive, Damped Trend Exponential Smoothing, Linear (Holt) Exponential Smoothing, Double (Brown) Exponential Smoothing, Random Walk with Drift, as well as various types of ARIMA and SARIMA. The accuracy of these methods was evaluated using essential metrics, namely Mean Squared Error (MSE), RootMean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The findings revealed that the ARIMA (5,0,4) model outperformed other models (Figure 1), demonstrating superior performance in predicting MMSNs among the traditional methods, with an MSE of 580.949, RMSE of 24.103, MAE of 17.19, and MAPE of 0.511.

To further enhance the accuracy of the predictions, a hybrid model was proposed. This model combined the traditional ARIMA (5,0,4) model with various machine learning methods, including multilayer perceptron, support vector machine, lazy algorithms, random forest, fast decision tree learner, M5 model tree algorithm, AR, RegbyDisc, J48, and their individual and combined approaches. Among these algorithms, the combination of AR with RegbyDisc based on J48 yielded the best results, highlighting its effectiveness

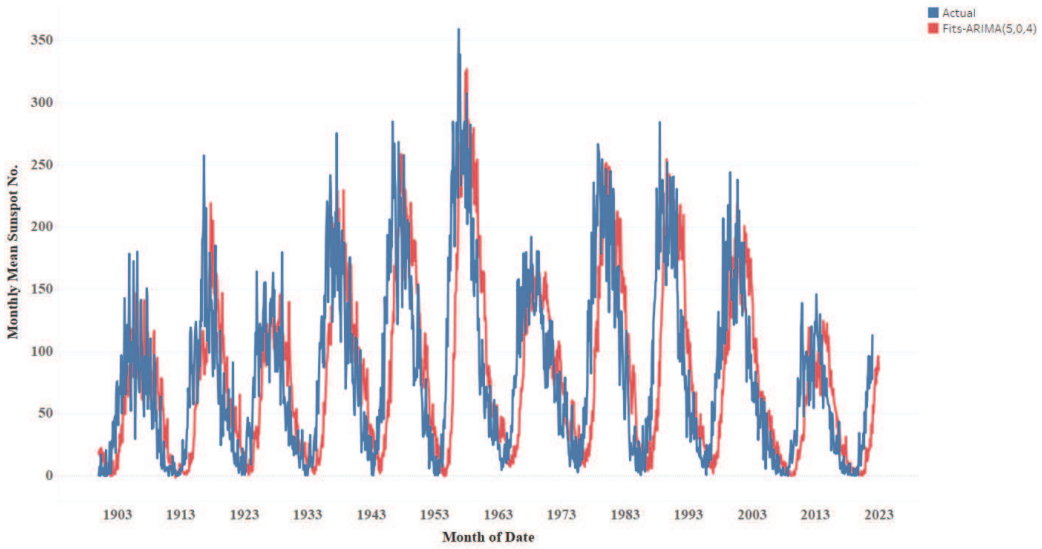

Figure 2: Actual and ARIMA(5,0,4) Predicted MMSNs

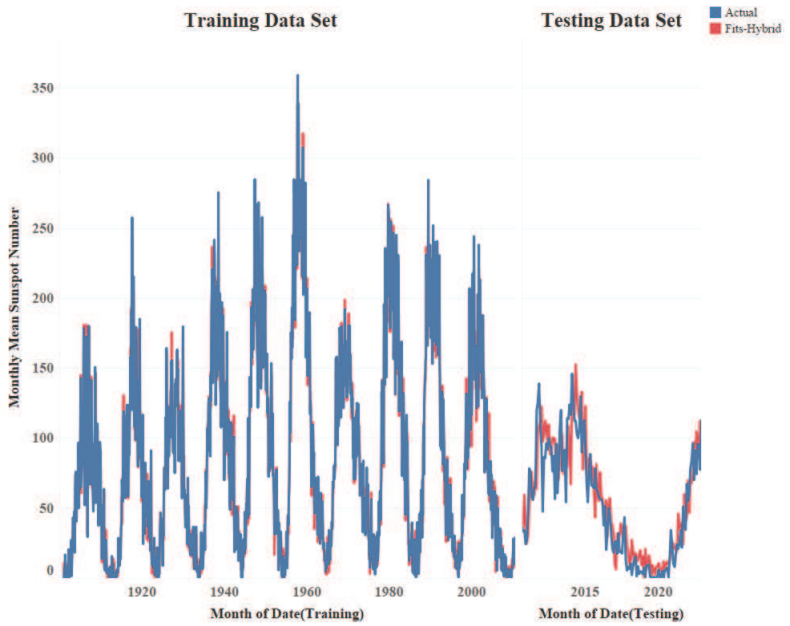

Figure 3: Hybrid Prediction of MMSNs: Training and Testing Comparison

in capturing the complex patterns within the sunspot time series data. This phase includes by incorporating an AR model that utilizes RegbyDisc based on J48. This approach is employed to effectively model the residuals derived from the ARIMA model. As the ARIMA model primarily captures linear patterns and may overlook nonlinear structures within the data, the residuals of the linear model can provide valuable insights into the nonlinearity present in the data.

Figure 3 illustrates a comparison between the actual and forecasted values using our proposed hybrid model for training and testing MMSNs. The hybrid model integrates the ARIMA and AR method by utilizing RegbyDisc based on J48. Notably, our proposed model outperforms the ARIMA model, demonstrating superior results. The hybrid model achieves the MSE of 114.653, RMSE of 10.708, MAE of 6.441, and MAPE of 0.438, further highlighting its improved performance. Leveraging our innovative hybrid approach, we have conducted solar predictions for the upcoming three years, revealing an anticipated peak SN of 164.63 for Solar Cycle 25 in December 2024 (Figure 4).

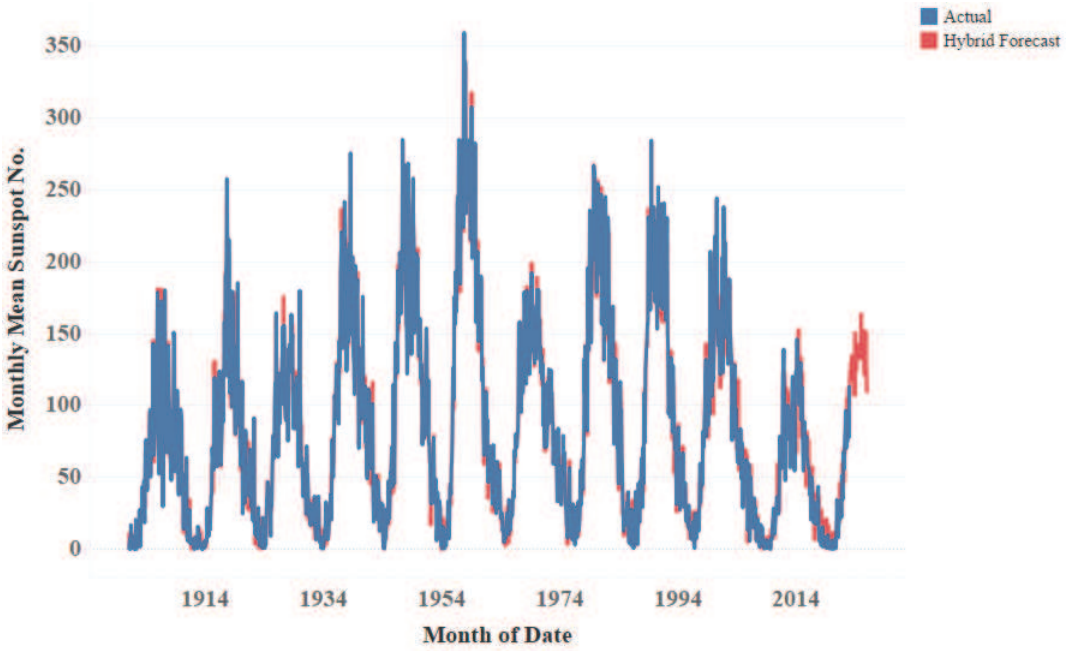

Figure 4: Hybrid MMSN forecasts: Jan 2023 - Dec 2025.

5 Conclusion

The study aimed to develop a reliable hybrid prediction model for sunspot time series data from January 1900 to December 2022 and utilized it for forecasting from January 2023 to December 2025. After a comprehensive analysis of forty-two predicting methods, the findings showed that the ARIMA(5,0,4) model demonstrated superior performance in predicting MMSNs compared to other traditional methods. To further enhance prediction accuracy, the study incorporated individuals and combinations of various popular machine learning algorithms with the ARIMA(5,0,4) model. The proposed hybrid model, ARIMA(5,0,4)-AR-RegbyDisc-J48, achieved markedly superior results, and we use it for MMSNs forecasting from January 2023 to December 2025.

References

[1] G. Ambika, K. P. Harikrishnan, "Methods of nonlinear time series analysis and applications: A review," Dynamics and Control of Energy Systems, (2020), 9-27.

[2] Yuanhui Fang,, Yanmei Cui, Xianzhi Ao, "Deep learning for automatic recognition of magnetic type in sunspot groups," Advances in Astronomy, 2019, (2019).

[3] Yuchen Dang, Ziqi Chen, Heng Li, Hai Shu, "A comparative study of non-deep learning, deep learning, and ensemble learning methods for sunspot number prediction," Applied Artificial Intelligence, **36,** no. 1, (2022), 2074129.

[4] Min Lei, Guang Meng, "Detecting nonlinearity of sunspot number," International Journal of Nonlinear Sciences and Numerical Simulation, **5**, no. 4, (2004), 321–326.

[5] Luis A. Aguirre, Christophe Letellier, Jean Maquet, "Forecasting the time series of sunspot numbers," Solar Physics, **249,** (2008), 103–120.

[6] Kristóf Petrovay, "Solar cycle prediction." Living Reviews in Solar Physics, **7,** no. 1, (2010), 1–59.

[7] Peter G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, **50,** (2003), 159–175.

[8] Indranil Bose, Xi Chen, "Hybrid models using unsupervised clustering for prediction of customer churn," Journal of Organizational Computing and Electronic Commerce, **19,** no. 2, (2009), 133–151.

[9] Wei Fan, "Prediction of Monetary Fund Based on ARIMA Model," Procedia Computer Science, **208,** (2022), 277–285.

[10] Jerome H. Friedman, Werner Stuetzle, "Projection pursuit regression," Journal of the American Statistical Association, **76,** no. 376, (1981), 817–823.

[11] Ian H. Witten, Eibe Frank, "Data mining: practical machine learning tools and techniques with Java implementations, " ACM Sigmod Record, **31,** no. 1, (2002), 76–77.

[12] Kevin Joshua Abela, Don Kristopher Angeles, Jan Raynier Delas Alas, Robert Joseph Tolentino, Miguel Alberto Gomez, "An automated malware detection system for Android using behavior-based analysis AMDA." International Journal of Cyber-Security and Digital Forensics,**2,** no. 2, (2013), 1–11.

[13] Ping-Feng Pai, Chih-Sheng Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting," Omega, **33,** no. 6, (2005), 497–505.

[14] Erasmo Cadenas, Wilfrido Rivera, "Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model," Renewable Energy, **35,** no. 12, (2010), 2732–2738.