Solar Physics (2024) 299:25
https://doi.org/10.1007/s11207-024-02270-6

RESEARCH

Data-Driven Forecasting of Sunspot Cycles: Pros and Cons of a Hybrid Approach

Qinglin Xu[1] · Rekha Jain[1] · Wei Xing[1]

Received: 12 October 2023 / Accepted: 6 February 2024

Abstract
Understanding the number of sunspots is crucial for comprehending the Sun's magnetic-activity cycle and its influence on space weather and the Earth. Recent advancements in machine learning have significantly improved the accuracy of time-series predictions, revealing a compelling approach for sunspot forecasts. Our work takes the pioneering work by proposing a hybrid forecasting approach that combines the Seasonal Autoregressive Integrated Moving Average (SARIMA) with machine-learning algorithms like Random Forest and Support Vector Machine, delivering high prediction accuracy. Despite its high accuracy, we highlight the need for caution in deploying machine-learning-based methods for sunspot-number prediction, demonstrated through a detailed case study with only three extra time stamps leading to a dramatic change. More specifically, when making a forecast of monthly averaged sunspot numbers from 2023–2043 based on data from 1749–2023, we found that the observations in June, July, and August 2023 have a significant impact on the forecast, particularly in the long term. Given the multiseasonal and nonstationary nature of the sunspot time series, we conclude that this kind of phenomenon cannot be simply captured by a pure data-driven model, which can be highly sensitive in the forecast in the long term, and requires a more comprehensive approach, possibly with a model that includes physics.

Keywords Sunspots · Solar cycle · Machine learning · Space weather

1. Introduction

The Sun's activity cycle is manifested in the number of sunspots on the surface of the Sun. It is monitored with great interest as its influence on Earth's atmosphere and communication channels is well documented (www.swpc.noaa.gov/). Sunspots are cooler regions on the Sun's surface that appear like dark spots in the observed images of the Sun. These spots also have very strong magnetic field compared to their surroundings. Sunspots can appear in a group or as an individual spot and can last for weeks.

Sunspots are believed to host many flaring activities in the Sun. When a flare erupts, energy equivalent to a billion megatons of TNT can be released in a few minutes. Hot erupting

✉ R. Jain
R.Jain@sheffield.ac.uk

[1] School of Mathematics and Statistics, University of Sheffield, Sheffield, S3 7RH, UK

Published online: 27 February 2024

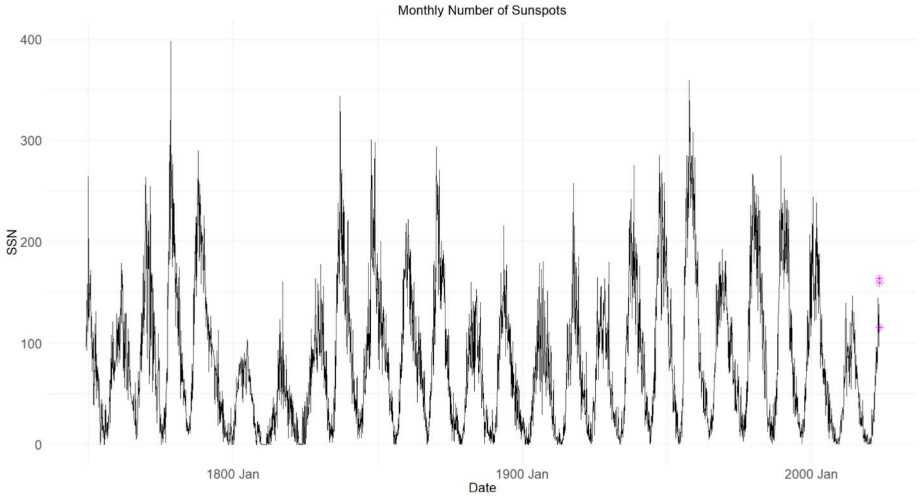

Figure 1 Sunspot count over time. The x-axis represents the years ranging from 1749 to May 2023, with a minimum scale of 1 month. The y-axis indicates the number of sunspots observed. Source: www.sidc.be/silso.

plasma released from the Sun traveling towards the Earth can interact with the Earth's magnetic field and cause geomagnetic storms (see www.weather.gov/fsd/sunspots). When there are a large number of sunspots, the Earth's geomagnetic activity is found to be increased and as a result the chances of damage to radio transmissions and satellite communications with impact on the navigation of planes, ships, and satellites is expected to increase.

Auroral activity on the Earth, a spectacular phenomenon in the night sky (www.spaceweatherlive.com/en/auroral-activity.html) is also seen more often during the periods of increased solar eruptions, but it is the adverse effects of solar activity on human technology, particularly from the large eruptions that reach the Earth, which is a big concern and how to minimize such adversities has posed a great challenge. It has been realized that predicting solar cycles is conducive to minimizing the harmful effects of magnetic-storm activity. Understanding of the formation and evolution of sunspots and the prediction of sunspot numbers on the Sun's surface is a critical area of research in solar physics.

In 1848, the Swiss astronomer Wolf suggested the relative number of sunspots (Wolf number) to represent Sunspot activity. This relative number, denoted by R, is defined as $R = K(s + 10g)$, where s is the number of individual spots and g is the number of sunspot groups. The parameter K is a constant determined by the quality of the observation. The daily numbers of sunspots have been recorded almost uninterrupted since 1610 AD with monthly average numbers, from 1749 AD onwards. From the long-term relative number of sunspots records, the average relative number of sunspots obviously shows a periodicity of about 11 years with the shortest being 9 years and the longest 13.6 years. During a solar cycle, the total number of sunspots rises rapidly, reaches a peak and then falls gradually. The highest number of sunspots in a cycle is referred to as the solar maximum of that cycle and the lowest number as the solar minimum of that cycle. Figure 1 shows the plot of the monthly mean total sunspot number (SSN) dating back from the year 1749. This data of SSN is taken from the SILSO website (www.sidc.be/silso/datafiles). The current sunspot cycle is referred to as Cycle 25. The three asterisks represent the unusually high number of sunspots in the months of June, July, and August in 2023.

It is clear from Figure 1 that there exists a nonlinear component in SSN in addition to a linear relationship. It is important to capture the essential ingredients of this nonlinear relationship in SSN to accurately predict them. Therefore, the methods used, in addition to the traditional prediction model, the Autoregressive Integrated Moving Average (ARIMA) model, vary greatly. For example, Yu et al. (2012) used Bayesian methods to predict the amplitude of Cycle 25 that was consistent with the predictions published by NOAA. Machine-learning algorithms have also been used in recent years. For example, Pala and Atici (2019) used Long-Short-Term Memory (LSTM) in deep learning for the prediction and found results that differ significantly from the NOAA predictions. The emergence of hybrid methods (Panigrahi et al., 2021) may be powerful enough to effectively capture the linear and nonlinear components in the time series of SSN. Appropriately motivated, we develop one such hybrid approach to predict the distribution of SSN in the next two solar cycles, given the number of sunspots available from January 1749 to August 2023. Our main aim is to investigate if there are major differences in the overall distribution of SSN in the next two cycles compared to the Sunspot Cycle 24.

Since the number of sunspots from June–August 2023 are unusually high, we carry out the same analysis on two separate datasets: (a) January 1749–May 2023 and (b) January 1749–August 2023. The paper is organized as follows. Sections 2–4 describe the dataset (a). In Section 2, we briefly describe the preliminary data-analysis methods. Section 3 provides details of the methods used in our hybrid approach. Section 4 reports the forecast results of the remaining portion of the sunspot Cycle 25 and the full sunspot Cycle 26 by three different methods and their comparisons. Section 5 shows the forecast results for dataset (b). Finally, a brief discussion and conclusion is given in Section 6.

2. Preliminary Data Analysis

The frequency distribution of the observed sunspot numbers for dataset (a), i.e., for January 1749–May 2023, is plotted in Figure 2. It is apparent that this distribution is right skewed. In order to minimize the effect of such a skewed distribution, we apply a power-law transformation. As an example, we carry out separate analyses for two different power-law indices: $\lambda = 0.5$ and 0.38. All the following figures are shown for $\lambda = 0.5$, referred to as *Square Root Transformation*, but we will compare the final results for both these values of indices.

The resulting frequency distribution for $\lambda = 0.5$ is shown in Figure 3. The transformed set of sunspot numbers will be referred to as SRSSN hereafter. We will then transform SRSSN to SSN for the final predictions.

2.1. Trend Forecast Using the SARIMA Model

It is well known that the Seasonal and Trend Decomposition using *locally estimated scatterplot smoothing* method referred to as the STL method, is very versatile and robust to decompose a time series such as the SRSSN. This method was proposed by Cleveland, Cleveland, and Terpenning (1990). Figure 4 shows this decomposition: from top to bottom panels, the SRSSN time series, trend component, seasonality with two different Periods (12 and 130), and residue are shown, respectively. As this is a monthly averaged data, the period 12 represents one year, and the period 130 represents approximately 130 months, which we consider as around 11 years. The term "remainder" refers to the remaining data after excluding the 12th and 130th periods, i.e., the data that persists after removing these possible

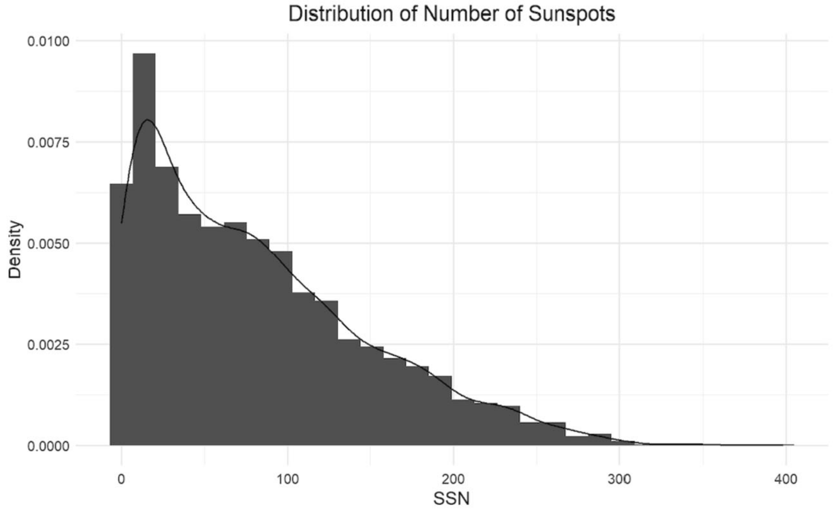

Figure 2 Distribution of the number of sunspots (SSN) as per Figure 1. The x-axis represents the range of SSN, ranging from 0 to the maximum SSN value. The y-axis represents the density of SSN, ranging from 0 to 0.01. The distribution forms a right-skewed pattern, where the mode is less than the median, which, in turn, is less than the mean.

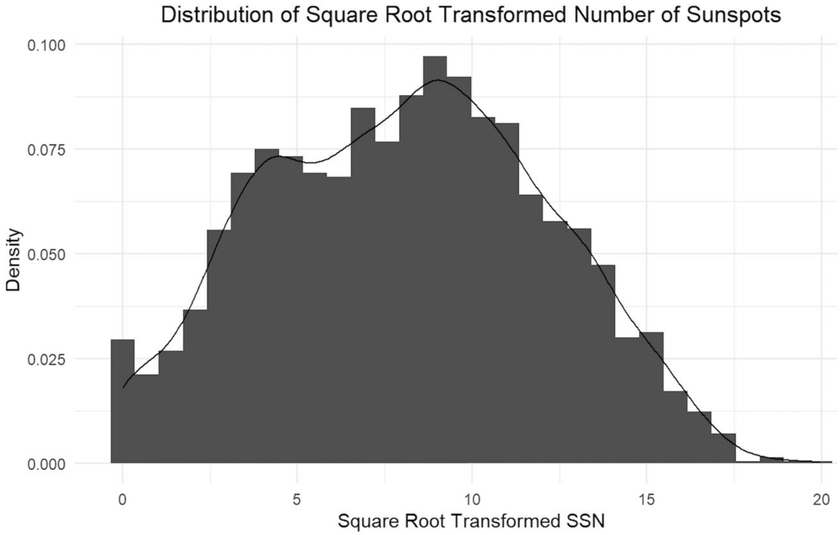

Figure 3 Square root transformed set of sunspot numbers (referred to as SRSSN). The x-axis represents SRSSN data ranging from 0 to 20, while the y-axis represents density from 0 to 0.1. The bimodal distribution exhibits a more symmetrical distribution than the right-skewed pattern of Figure 1.

periodicities. In this context, “remainder” does not refer to residuals we use for prediction in the later section.

It can be noted from Figure 4 that the periodicity is more evident for Period = 130, indicating that the sunspot data indeed follows a cycle of approximately 11 years. Additionally,

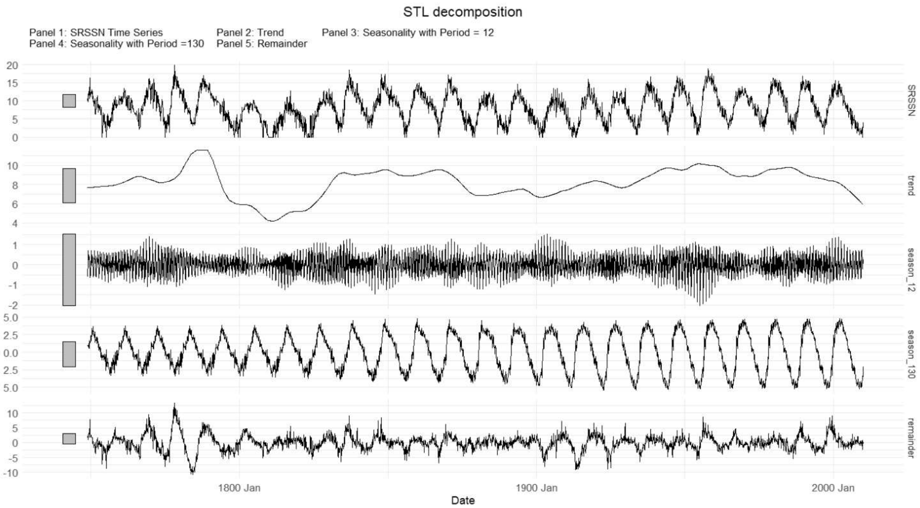

Figure 4 STL decomposition of SRSSN. The x-axis represents the date range from January 1749 to May 2023, which applies to all five panels. The panels, arranged from top to bottom, show the SRSSN time series, trend component, seasonal terms, and remainder, respectively.

there is also some possible periodicity in the sunspot data for Period $= 12$, suggesting that in addition to the 11-year cycle, an annual pattern may also exist. When making predictions, it is important to consider both these periodicities and hence the use of SARIMA in forecasting.

As a first step in our forecasting procedure, we divide the dataset into two sets: The training set and the test set. We use the training set to make predictions. This predicted set is then fitted onto the test set to obtain the trend-fit model. We choose data up to the year 2009 as the training set and the subsequent data as the test set. Figure 5 shows the predicted set with the blue color plotted on top of the original test dataset shown with a black curve. The predicted test set mostly overlaps with the original data. This indicates the feasibility of the SARIMA method.

To identify the optimal SARIMA $[p, d, q][P, D, Q]$, we fitted the parameters ranging from 0 to 1 for p, d, q, P, D, Q with various seasonal frequencies from 120 (10 years) to 144 (12 years) on the training data to predict the test data. A total of 1600 sets of results were obtained and compared based on their Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The result with the minimum RMSE and MAE was selected to forecast the SSNs for the next 20 years. The results showed that a frequency of 140, $p = 1$, $d = 1$, $q = 0$, $P = 1$, $D = 1$, $Q = 0$ yielded the best-fitting performance among the 1600 sets. In Figure 6, we show the trend set for the entire SRSSN. It shows the original transformed data (SRSSN) represented by the gray line. The blue line represents the trend obtained by fitting the entire dataset using SARIMA. Subtracting this trend from the actual SRSSN data yield the residuals.

We find that the residuals are almost impossible to predict using SARIMA. However, the trend-fit model for the SARIMA method can be used for forecasting the trend set for future sunspot cycles. In Figure 7, we show the trend forecast for the forthcoming 20 years from June 2023 to May 2043. The top panel (a) shows the complete dataset, while panel (b) is a zoomed-in view for more clarity.

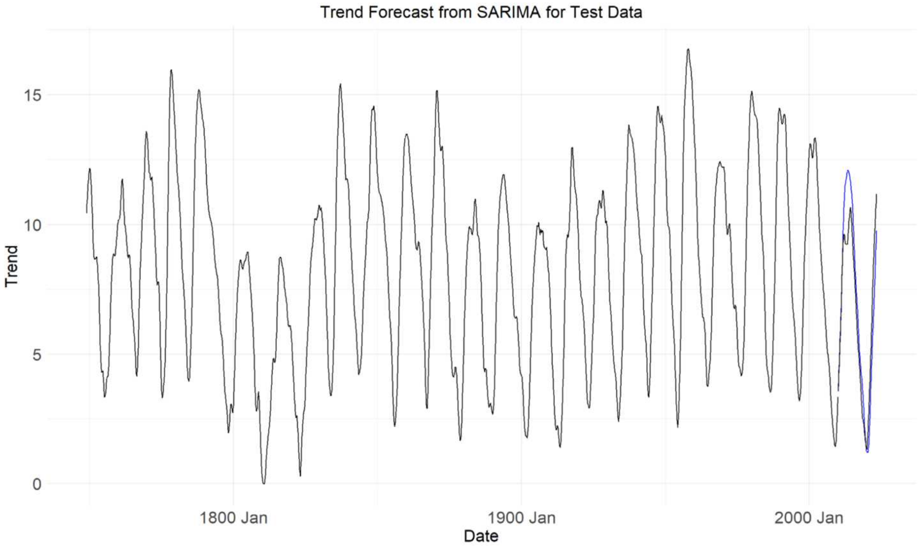

Figure 5 The trend forecast of the SRSSN using the SARIMA method for the test data. The x-axis represents the period of the training set from January 1749 to December 2009, and the test set from January 2010 to May 2023. The y-axis represents the SRSSN data ranging from 0 to 20. The original SRSSN data is shown by the *black line*, while the *blue line* represents the predicted data for the test set.

3. Methods Used for the Hybrid Approach

In the previous subsection, we discussed the trend dataset using the SARIMA method. In this section, we consider three machine-learning algorithms, the support vector machine (SVM), random forests (RF), and XGBoost to fit the residuals of the test set. We refer to the combined use of SARIMA and the machine-learning algorithm as the Hybrid approach/model.

Recall that the residual series were obtained by subtracting the actual data from the fitted trend data. Additionally, predictive variables were constructed by combining the mean, square median, maximum, minimum, and standard deviation of the residual series across different months and quarters. Principal Component Regression (PCR) and Partial Least Squares (PLS) were applied to reduce the dimensionality of the predictions, and the performance of PCR and PLS was measured by their R-squared value. It was found that PLS (R-squared = 0.0036) performed better than PCR (R-squared = 0.0011). In addition to the Hybrid model, we also consider another advanced method, Seasonal and Trend decomposition using Loess (STL) and compare its performance with the hybrid method.

3.1. Model Evaluation

The performance of all hybrid models can be judged more intuitively by comparing Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). In Figure 8, we illustrate these quantities for all the machine-learning algorithms that we examined, namely, SVM, RF, XGBoost, and their simple average for the case of $\lambda = 0.5$. Careful comparison considering MAE and RMSE suggests that RF outperforms the others. However, we will select two different combinations, SARIMA + RF and SARIMA + SVM to forecast the maxima and minima of the next two sunspot cycles given the number of sunspots available from January 1749 to May 2023 for the square-root-transformed dataset.

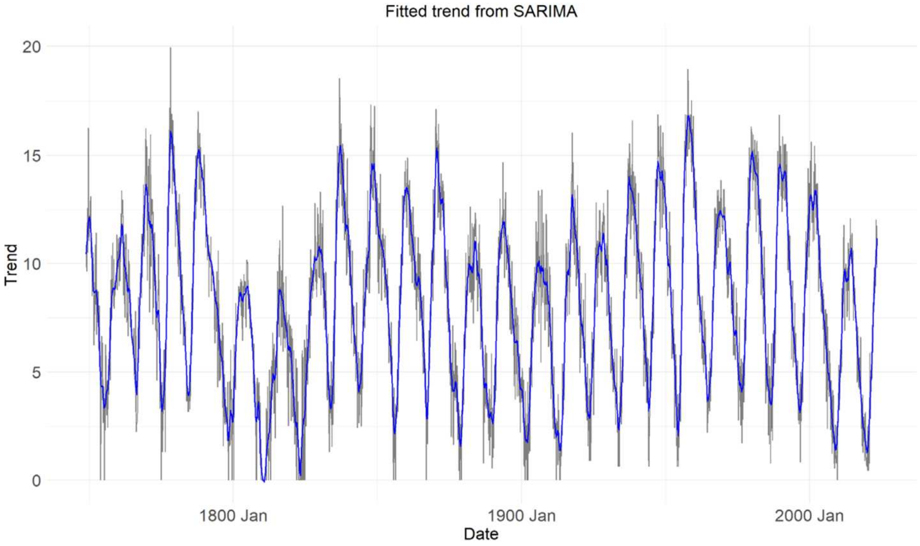

Figure 6 Fitted SRSSN trend value of the SARIMA method. The x-axis denotes the date from January 1749 to May 2023. The y-axis denotes the SRSSN data with the range from 0 to 20. The *gray line* represents the actual data and the *blue line* represents the fitted trend set. The *blue line* almost overlaps the *gray line* with a later starting point.

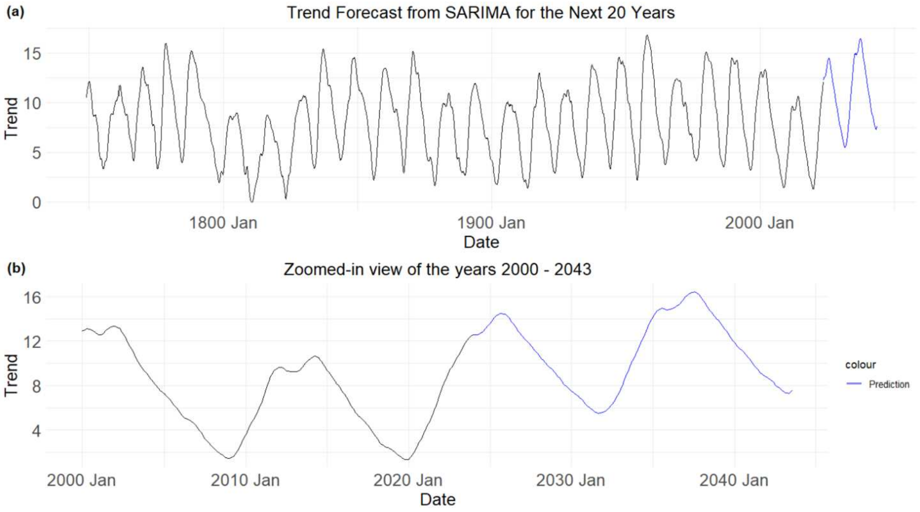

Figure 7 (a): The trend forecast of the SARIMA method for the forthcoming 20 years of SRSSN. The x-axis represents the period from January 1749 to May 2023 for the original SRSSN data, and from June 2023 to November 2042 for the predicted trend. The y-axis represents the SRSSN data ranging from 0 to 20. The *black line* indicates the original trend set, while the *blue line* represents the forecasted trend set. The panel (b) is the same as panel (a) except zoomed into the years from 2000 to 2042.

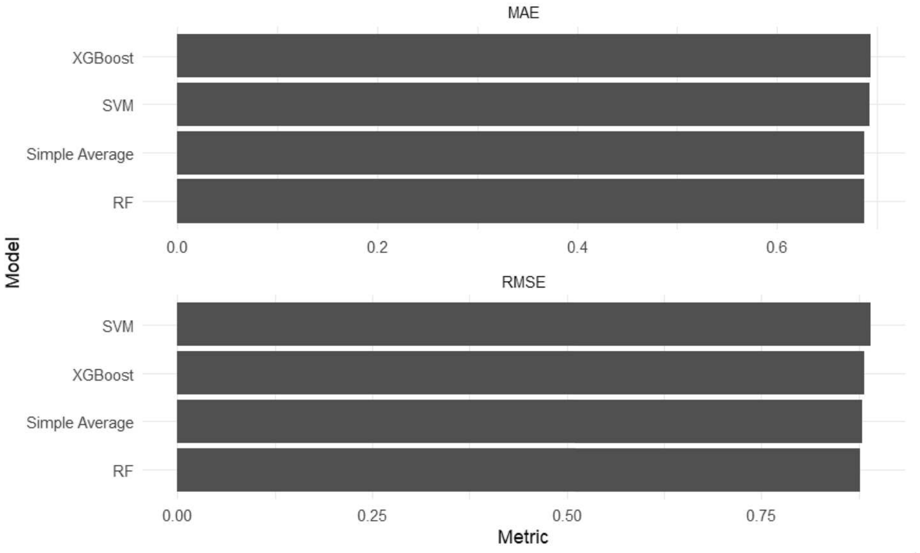

Figure 8 The performance comparison of machine-learning algorithms, SVM, RF, XGBoost, and their simple average. MAE and RMSE shown with the *shaded bars* for the four algorithms suggest that their performance is quite similar with RF performing slightly better.

4. Prediction of Sunspot Cycles for Dataset (a)

In Figure 9, we show the predicted SRSSN using the SARIMA + RF hybrid model for dataset (a). The top panel displays both, the original (black) and the predicted (blue) dataset with the bottom panel displaying the zoomed-in view of the predicted dataset. Despite the slight difference in the durations of the cycles, the qualitative behavior of the two predicted cycles suggests that they are not going to be significantly different. It would now be interesting to compare the actual number of sunspots as a function of time from this method with the standard STL method. Figure 10 compares the forecast of SSN from SARIMA + RF (blue) and STL (red). The predicted SSNs show a slight delay in maxima for the SARIMA + RF hybrid model.

Similar to Figure 9, Figure 11 displays the predicted SRSSN as a function of time from the SARIMA + SVM hybrid model. The black curves are the original observational data and the blue curves are our forecast. The top panel shows the full dataset and the bottom panel is a zoomed-in view from year 2000 to 2043, which clearly shows two similar peaks, Sunspot Cycle 25 (SC25) and Sunspot Cycle 26 (SC26). Both cycles also show a gradual decay phase then the rise phase.

Figure 12 compares the predicted SSN from the SARIMA + SVM model (blue curves) with the STL (red curves) method. Once again, similar to the predictions by SARIMA + RF shown in Figure 10, the maxima and minima of the two curves suggest a slight delay even for SARIMA + SVM compared to STL with SC26 not too dissimilar to SC25 or perhaps slightly weaker than SC25.

We now discuss the predicted SSN from SARIMA + RF in more detail. We compare the predicted SSN obtained using the hybrid SARIMA + RF model for two different types of power-law transformations ($\lambda = 0.38$ and 0.5) in Table 1. Although there is a disparity in the exact number of SSN and the month of maximum, both columns show the same 2025

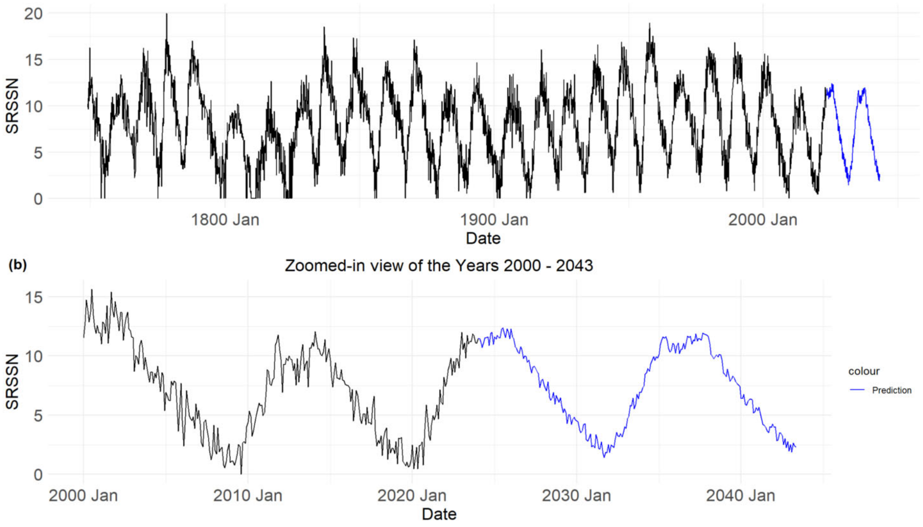

Figure 9 The forecast of SRSSN for the next 20 years with the SARIMA + RF model. The x-axis represents the time period from January 1749 to May 2023 (original data) and from June 2023 to November 2042 (prediction period). The y-axis represents the predicted SRSSN data ranging from 0 to 20. The *black line* depicts the actual data, while the *blue line* represents the predicted data with SARIMA + RF. Panel (b) displays the zoomed-in view of the predicted curves along with a portion of the original data.

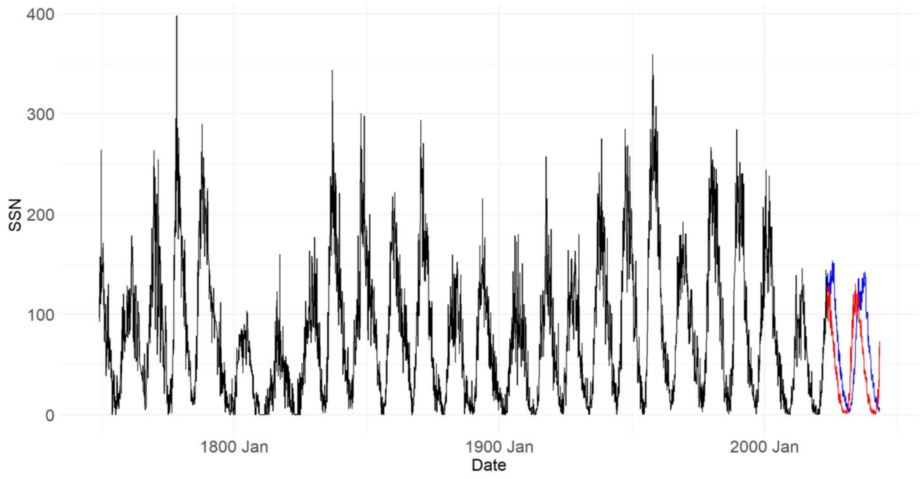

Figure 10 Comparison of SSN for the next 20 years obtained with the SARIMA + RF model and STL. The x-axis represents the time period from January 1749 to May 2023 (original data) and from June 2023 to November 2042 (prediction period). The y-axis represents the predicted SSN data ranging from 0 to 400. The *black line* depicts the actual data, while the *blue curves* represent the predicted data with SARIMA + RF and the *red curves* are for STL with multiple seasonality included.

year for the maximum of SC25. For the Sunspot Cycle 26, the maximum is predicted to be in September 2036 for $\lambda = 0.38$ and September 2037 for $\lambda = 0.5$. The minimum for SC27 is forecast as 2042 ± 2 months.

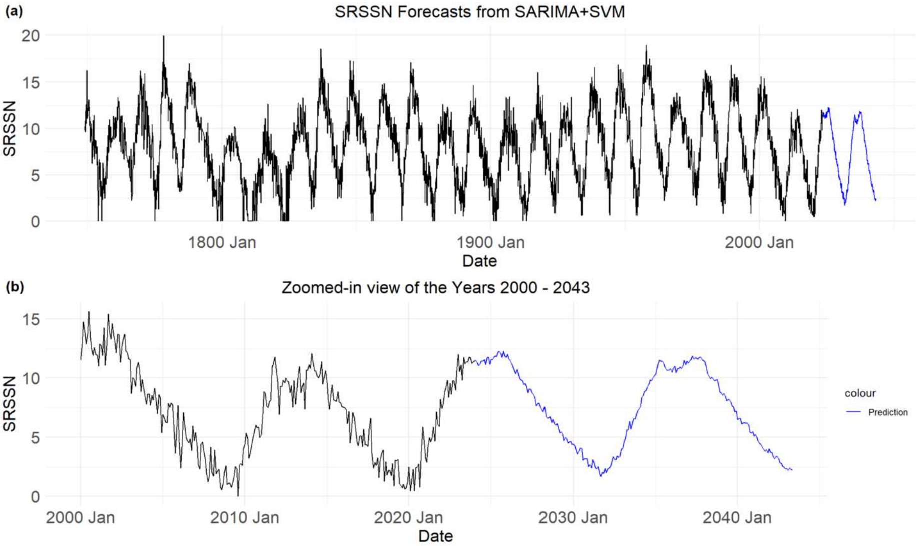

Figure 11 Comparison of SSN for the next 20 years obtained with SARIMA + SVM and STL. The x-axis represents the time period from January 1749 to May 2023 (original data) and from June 2023 to November 2042 (prediction period). The y-axis represents the predicted SSN data ranging from 0 to 400. The *black line* depicts the actual data, while the *red line* represents the predicted data with SARIMA + SVM and *blue lines* are for STL with multiple seasonality included.

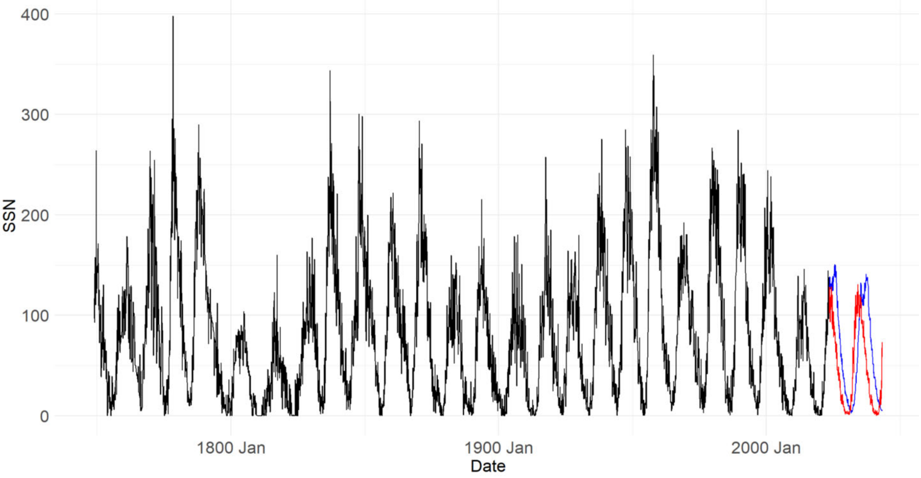

Figure 12 Comparison of SSN for the next 20 years obtained with SARIMA + SVM and STL. The x-axis represents the time period from January 1749 to May 2023 (original data) and from June 2023 to November 2042 (prediction period). The y-axis represents the predicted SSN data ranging from 0 to 400. The *black line* depicts the actual data, while the *red line* represents the predicted data with SARIMA + SVM and *blue lines* are for STL with multiple seasonality included.

The future solar cycles are predicted to be quite similar for SSN even when two different values of lambda were used to transform the original dataset (a).

Table 1 Forecast of SSN: comparison between λ = 0.38 and 0.5 for the SARIMA + RF Hybrid model for dataset (a).

λ	0.38	0.5
Sunspot Cycle 25		
Maximum date (SSN)	October 2025 (261)	July 2025 (154)
Sunspot Cycle 26		
Minimum date (SSN)	September 2031 (11)	September 2031 (2)
Maximum date (SSN)	September 2036 (205)	September 2037 (141)
Sunspot Cycle 27		
Minimum date (SSN)	December 2042 (24)	February 2043 (3) but not at the beginning of the cycle

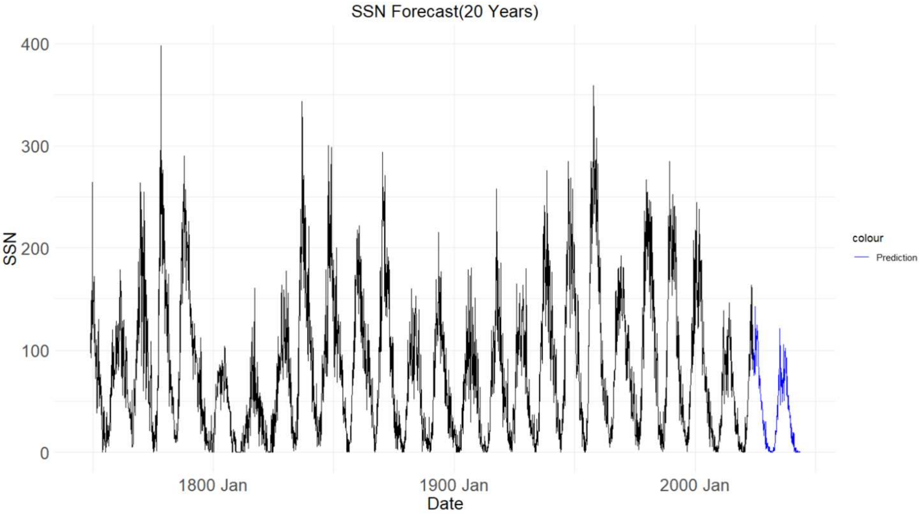

Figure 13 Predicted SSN shown in *blue* with SARIMA for the next 20 years using the dataset (b) without splitting the trend and residuals.

5. Prediction of Sunspot Cycles for Dataset (b)

We now carry out a similar analysis for dataset (b), i.e., we now use the monthly averaged sunspot number between January 1749–August 2023 and predict SSN for the next 20 years. As mentioned earlier, SARIMA alone is not appropriate for predicting residuals. This can be seen by comparing Figures 13 and 14.

In Figure 13, SARIMA(1,1,0)(1,1,0), frequency = 140 is used directly on SRSSN without splitting trend and residuals for forecasting. The final result is the prediction after reversing the transformation.

This can be compared with Figure 14 where the STL decomposition is first used on SRSSN to extract the trend component. The forecasted trend is obtained from SARIMA(1,1, 0)(1,1,0), frequency = 140. Subtracting the fitted trend from the SRSSN data then yields the residuals. We then forecast the residuals using the SARIMA method, combine it with the

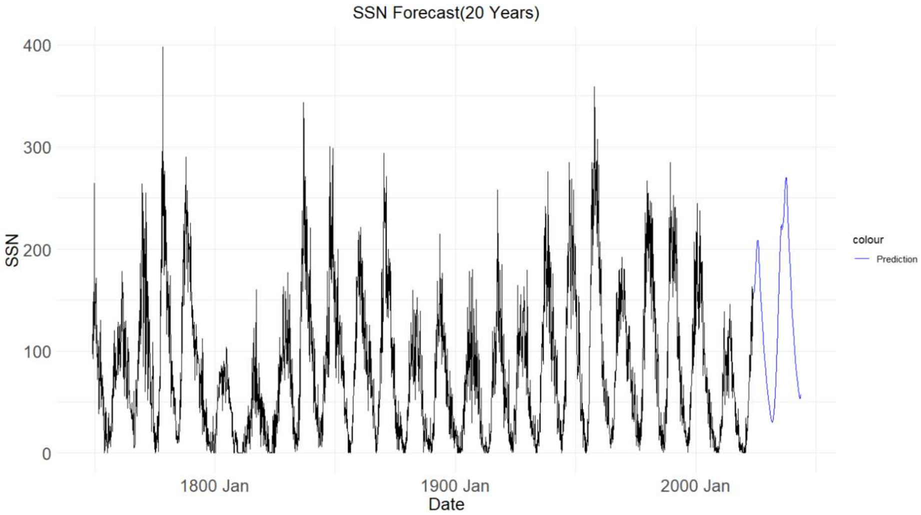

Figure 14 Predicted SSN shown in *blue* with SARIMA for the next 20 years using the dataset (b). The STL decomposition was used to separate the trend and residuals and SARIMA was used to forecast each before combining them.

predicted trend, and finally reverse the transformation to obtain the final prediction for SSN. The predicted residuals appear to be zero. Thus, we conclude that the predictive performance of SARIMA on the residuals is relatively poor for this time series and proceed with the Hybrid model.

Figure 15 first shows the SRSSN forecast using the SARIMA + RF model, while Figure 16 compares them with STL. It is obvious from Figures 15 and 16 that the predicted SRSSN for dataset (b) are quite high when compared with the corresponding ones for dataset (a). Similarly, when SVM is used in place of RF in the Hybrid model, the same high numbers for SRSSN and SSN can be seen in Figures 17 and 18. Thus, the number of SSNs is substantially more by just including the last three additional datapoints and we expect the maxima and minima to be quite different for the two datasets for the same hybrid models. As an example, we compare the minimum and maximum obtained from the SARIMA + RF hybrid model for both datasets (a) and (b).

Table 2 compares the forecasts for SSN with the STL method and with the SARIMA + RF Hybrid model. For Cycle 25, both methods STL and RF predict the year of maxima as 2024 and 2025, respectively. There is a significant difference in the predicted maximum and minimum years for Cycle 26 and 27.

6. Discussion and Conclusion

The number of observed sunspots with time show multiple periodicities. The study of these periodicities is important for understanding the physical mechanisms responsible for solar magnetic activity and for forecasting the maxima of the next sunspot cycles with the aim of minimizing the adverse effects of solar activity on space weather. Efforts are being continued to understand the observational data by including physical effects in the numerical simulations of solar dynamo (see, for example, Charbonneau, 2020 and references therein) but in recent times, there is also a focus on the use of machine-learning algorithms solely

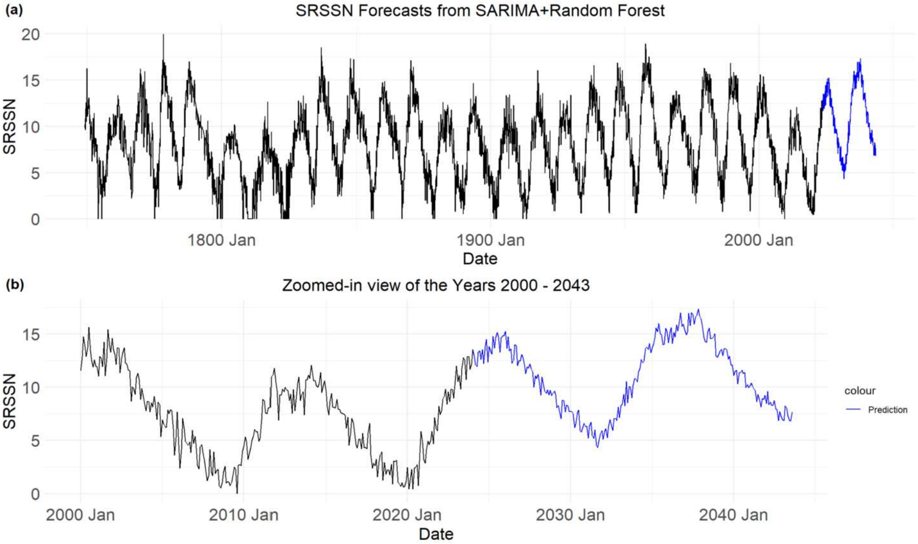

Figure 15 The forecast of SRSSN from dataset (b) for the next 20 years with the SARIMA + RF model. The x-axis represents the time period from January 1749 to August 2023 (original data) and from September 2023 to November 2042 (prediction period). The y-axis represents the predicted SRSSN data ranging from 0 to 20. The *black line* depicts the actual data, while the *blue line* represents the predicted data with SARIMA + RF. Panel (b) displays the zoomed-in view of the predicted curves along with a portion of the original data.

Table 2 Forecast of SSN comparison between STL and the SARIMA + RF Hybrid model ($\lambda = 0.5$) for the two datasets.

Method	STL	SARIMA + RF for dataset (a)	SARIMA + RF for dataset (b)
Sunspot Cycle 25			
Maximum date (SSN)	April 2024 (128)	July 2025 (154)	January 2026 (232)
Sunspot Cycle 26			
Minimum date (SSN)	October 2030 (1)	September 2031 (2)	September 2031 (19)
Maximum date (SSN)	March 2034 (131)	September 2037 (141)	November 2037 (301)
Sunspot Cycle 27			
Minimum date (SSN)	February 2041 (1)	February 2043 (3) but not at the beginning of the cycle	October 2042 (47)

with the aim to forecast the sunspot cycles using the long time series of observed sunspots (see, for example, Pala and Atici, 2019; Panigrahi et al., 2021). These later methods do not involve the physics of sunspots but instead directly detect the natural periodic patterns in the observational data.

In the present study, we have used a hybrid model consisting of statistical analysis and a machine-learning algorithm to predict the maxima of sunspot Cycles 25 and 26 and the minima of SC 26 and SC27. We selected two appropriate machine-learning algorithms, Random Forest (RF) and Support vector machine (SVM), to combine with SARIMA to

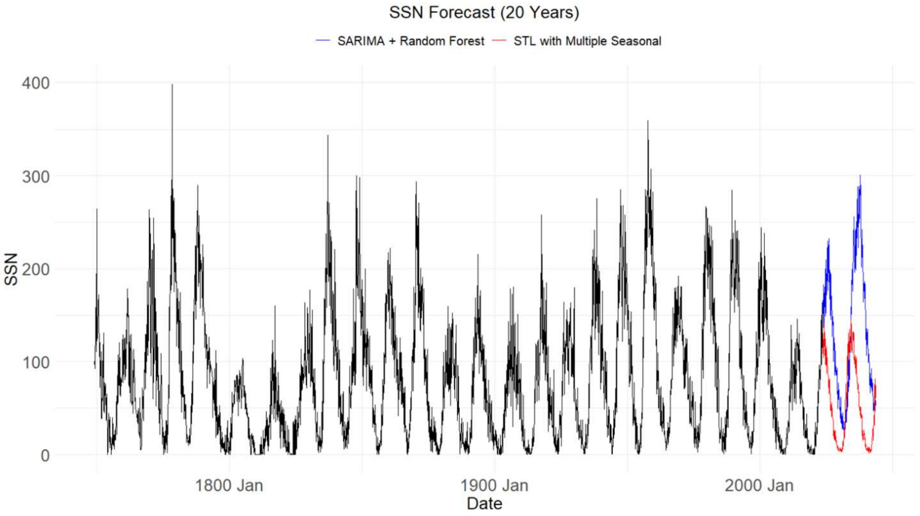

Figure 16 Comparison of SSN for the next 20 years obtained with the SARIMA + RF model and STL from dataset (b). The x-axis represents the time period from January 1749 to August 2023 (original data) and from September 2023 to November 2042 (prediction period). The y-axis represents the predicted SSN data ranging from 0 to 400. The *black line* depicts the actual data, while the *blue curves* represent the predicted data with SARIMA + RF and the *red curves* are for STL with multiple seasonality included.

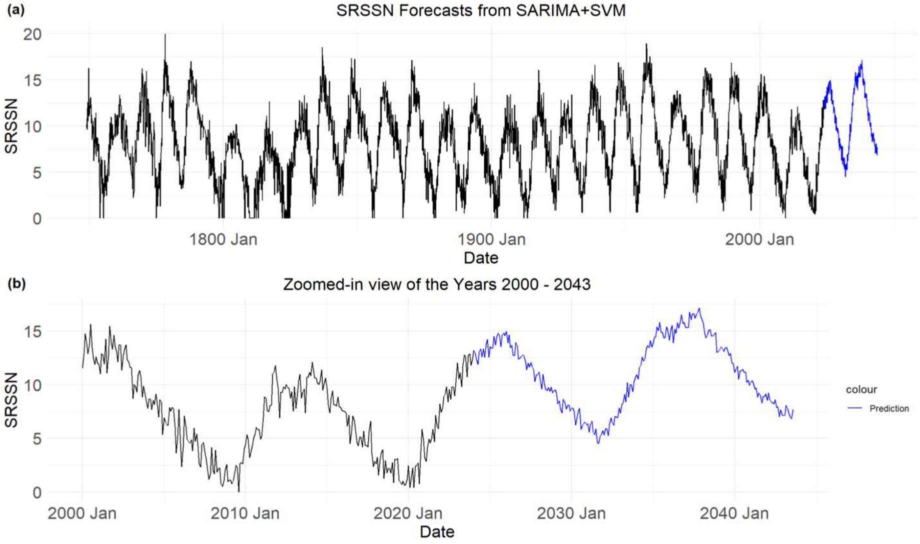

Figure 17 Same as Figure 11 but for dataset (b).

forecast the trend and residual set of future sunspot cycles. We divided the observed sunspot numbers into two sets: (a) January 1749–May 2023 and (b) January 1749–August 2023. A similar analysis was carried out for each dataset.

In our study, first the observed monthly averaged sunspot numbers (SSN) obtained from www.sidc.be/silso, are transformed to minimize the right-skewed pattern of the original

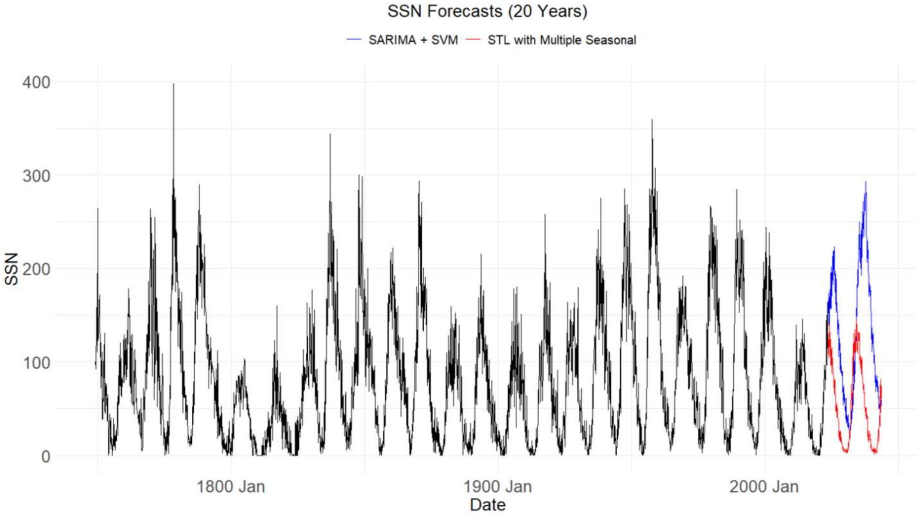

Figure 18 Same as Figure 12 but for dataset (b).

data by using a power-law transformation. Secondly, the resulting distribution is then decomposed to extract multiseasonality. Noting that the SARIMA method alone cannot be used for forecasting residuals, we explored machine-learning algorithms. Finally, in order to forecast SSN over the next 20 years, we then developed the code for the Hybrid model of SARIMA + RF and applied this to the original data. In addition to this combination the other machine-learning algorithms SVM, XGBoost, and their simple average were also examined for predicting the residuals but RF was found to yield slightly better performance; RF and SVM were each used for the prediction of residuals with SARIMA. We also explored the forecast using the STL method and its final results compared with the two developed Hybrid models.

To summarize, the prediction of sunspot cycles based on our Hybrid model SARIMA + RF applied to dataset (a) suggests that:

(1) SC25 will be stronger than SC24.
(2) The mean sunspot number at the maximum is sensitive to the choice of statistical parameters, but is expected to be greater than 150 for SC25.
(3) SC25 will reach a maximum in the year 2025 and SC26 in the year 2037.
(4) The minimum for SC26 is predicted to be in the year 2031 and is going to be slightly weaker than SC25 from dataset (a). However, it is predicted to be much stronger compared to SC25 from dataset (b).
(5) The STL method shows reduced SSN for each cycle compared to the Hybrid model.

The conclusion from our Hybrid approach that SC25 will be moderate and only slightly stronger than SC24 is not fully consistent with other recent model-based estimations (e.g., Javariah, 2017) that predict SC25 to be weaker in strength than SC24. Similarly, for Hathaway and Upton (2016) and Cameron, Jiang, and Schussler (2016), both studies were based on surface flux-transport code, predicted SC25 to be slightly weaker than SC24. Based on the spectral analysis of the three distinct series of the observed sunspots, Kane (2007) predicted the maximum sunspot numbers to be about 119 for Cycle 25 during years 2022–2023 and about 118 for Cycle 26 during years 2032–2034. However, a solar dynamo index prediction by Pesnell and Schatten (2018) estimated a sunspot number of 135 ± 25 for Solar

Cycle 25 making it comparable to Solar Cycle 24. According to them, the estimated peak for Cycle 25 is expected near 2025 ± 1.5 yr. Our estimates suggest higher numbers than these and at least two years later for the maxima for these cycles. The previous hybrid model of Panigrahi et al. (2021) suggested the maximum SSN to be 94.33 in April 2026, whereas Pala and Atici (2019) had predicted the maximum of SC25 in July 2022 with SSN = 167.3. Clearly, the Sunspot Cycle 25 is still rising with SSN, with the maximum SSN, so far, above 150.

Our approach is based extensively on first processing the actual observational data for extracting the seasonality and then using the trend from all the sunspot cycles prior to 2010 to fit the trend of the subsequent sunspot cycles (up to May 2023 in dataset (a) and up to August 2023 in dataset (b)). This trend-fit model is then used to forecast the trend of the future sunspot cycles. The residuals are obtained by subtracting the trend from the original data. An appropriate machine-learning algorithm is then applied to best fit the residuals that are then combined with the trend-fit model for forecasting the future sunspot numbers. We conclude from dataset (a) that Solar Cycles 25 and 26 are not going to be significantly different, with sunspot numbers slightly fewer in SC26. On the other hand, dataset (b) suggests that Solar Cycle 26 will be stronger than SC25.

The use of the Hybrid model has been advantageous in the sense that it can forecast residuals that was almost impossible with SARIMA alone. However, we have also shown that the same Hybrid model can give a very different outcome with a small increment in the input data. It should be noted that the physics hidden behind the formation and decay of sunspots has strong links with the sunspot activities and its ∼11-year periodicities on the surface of the Sun. Magnetic fields play a key role in the formation and decay of sunspots and since turbulent processes are responsible for the generation and transport of magnetic fields at various dynamic scales in the Sun, use of empirical models based on statistical relations of long-term observations like the one used in the present study is unlikely to provide a complete understanding of the sunspots themselves and why their numbers vary with time. The aim of the present analysis is not to rule out this link but it is merely to show the use of machine-learning algorithm for extracting information from the long time series of the observed sunspot numbers to forecast the qualitative behavior of the next few sunspot cycles and to show the sensitivity of statistical parameters in such empirical procedures. Ultimately, an approach that combines physics-based theoretical models (see, for example, Kitiashvili, 2016; Bhowmik et al., 2023) and the hybrid approach such as this one (see also, Panigrahi et al., 2021) to extract information from the long-term observational data may be needed for predicting more accurate sunspot numbers, their variation with time, and the physical mechanism that cause them.

Author contributions Q.X. analysed the data and prepared the figures. R.J. wrote the manuscript based, in part, on Xu (2022). X.W. provided some advice on the Machine-learning aspects. All authors reviewed the manuscript.

Data Availability The monthly average sunspot data is publicly available and is taken from SILSO World Data Center (https://www.sidc.be/SILSO/home).

Declarations

Competing interests The authors declare no competing interests.

References

Bhowmik, P., Jiang, J., Upton, L., et al.: 2023, *Space Sci. Rev.* **219**, 40.
Cameron, R.H., Jiang, J., Schussler, M.: 2016, *Astrophys. J.* **823**, L22.
Charbonneau, P.: 2020, *Living Rev. Solar Phys.* **17**, 1.
Cleveland, R.B., Cleveland, W.S., Terpenning, I.: 1990, *J. Off. Stat.* **6**(1), 3.
Hathaway, D.H., Upton, L.A.: 2016, *J. Geophys. Res. Space Phys.* **121**, 10,744.
Javariah, J.: 2017, *Solar Phys.* **292**, 11.
Kane, R.P.: 2007, *Solar Phys.* **246**, 487.
Kitiashvili, I.N.: 2016, *Astrophys. J.* **831**, 15.
NOAA, https://www.swpc.noaa.gov/news/solar-cycle-25-forecast-update.
Pala, Z., Atici, R.: 2019, *Solar Phys.* **294**(5), 50. DOI
Panigrahi, S., Pattanayak, R.M., Sethy, P.K., et al.: 2021, *Solar Phys.* **296**(1), 1.
Pesnell, W.D., Schatten, K.H.: 2018, *Solar Phys.* **293**, 112.
SILSO, World Data Center, Royal Observatory of Belgium - www.sidc.be/silso/datafiles
Xu, Q.: 2022, MSc. Statistics Dissertation.
Yu, Y., van Dyk, D.A., Kashyap, V.L., et al.: 2012, *Solar Phys.* **281**(2), 847.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.