AUA CSE, SPRING 2023, DS207 - Time Series Forecasting

TIME-SERIES FORECASTING OF MONTHLY SUN SPOT DATA

M. Davtyan, A. Baghdasaryan, L. Galstyan

MOTIVATION

- Sunspot data is important for understanding the sun's magnetic activity and its impact on Earth's climate and space weather.
- Daily observations of sunspot activity have been recorded by astronomers since 1818, and the data collected has been a significant area of research in solar physics.
- Time-series forecasting techniques involve using statistical and machine learning methods to predict future activity based on past observations and have been widely applied to analyze and predict sunspot activity.
- Time-series models can help identify patterns and trends in sunspot data and provide valuable forecasts for various stakeholders such as solar energy companies, climate scientists, space weather forecasters, and amateur astronomers.
- Daily sunspot data and time-series forecasting are essential tools for the study of the sun's behavior and its impact on Earth and the space environment.

LITERATURE REVIEW AND RESEARCH QUESTION

Research Question: How accurately can we forecast daily sunspot numbers aggregated monthly using time series forecasting models based on the available attributes in our dataset, and how does the variability of sunspot numbers change over time?

V. Sandeep Kumar. (2018). "Sunspot Activity Prediction with Neural Networks" (Kaggle):
The work by V. Sandeep Kumar explores the use of neural networks for sunspot activity prediction. The author utilizes a dataset from Kaggle and employs a multilayer perceptron (MLP) neural network model to forecast the monthly average sunspot activity. The model achieved promising results, demonstrating the potential of neural networks in sunspot prediction. The work provides valuable insights into the application of neural networks in solar physics research.

M. Ali. (2018). "Sunspot Forecasting Using Time Series Analysis" (GitHub):
M. Ali's work focuses on sunspot forecasting using time series analysis. The author presents a comprehensive analysis of the sunspot dataset and applies various time series techniques, including autoregressive integrated moving average (ARIMA) modeling and exponential smoothing methods. The study emphasizes the importance of data preprocessing and model selection in achieving accurate sunspot forecasts.

C. Y. Liu, S. Q. Cao, L. H. Lu, & X. Y. Zhang. (2013). "Forecasting Sunspot Cycles Using ARIMA Models" (Advances in Astronomy):

In this work, Liu et al. investigate the use of ARIMA models for sunspot cycle forecasting. The authors propose a methodology to predict the amplitude and timing of future sunspot cycles based on historical sunspot data. They demonstrate the effectiveness of ARIMA models in capturing the cyclic nature of sunspot activity and forecasting future cycles. The study contributes to the understanding of long-term sunspot behavior.

M. T. Ullah, M. Z. Uddin, & M. A. Matin. (2015). "A Hybrid ARIMA and Artificial Neural Network Model for Sunspot Forecasting" (International Journal of Scientific & Engineering Research):

Ullah et al. propose a hybrid model that combines ARIMA and artificial neural network (ANN) techniques for sunspot forecasting. The authors integrate the strengths of both approaches, utilizing ARIMA for short-term prediction and an ANN for long-term forecasting. The hybrid model demonstrates improved accuracy compared to individual ARIMA or ANN models. This study highlights the potential benefits of combining different forecasting methods for sunspot activity prediction.

DATA

Data Description:

The daily sunspot data set contains observations of the number of sunspots on the surface of the sun from the years 1818 to 2019. The data was collected by the Royal Observatory of Belgium, including daily sunspot counts and monthly and yearly averages. The data set contains 73,583 observations, with each observation corresponding to a single day. The data set can be used to study the patterns and trends in sunspot activity over time and to make predictions about future activity using time-series forecasting techniques. The data set is made available on Kaggle, a platform for sharing and discovering data sets for use in research and analysis.

Kaggle Link

COLUMN 1-3:

Gregorian calendar date
- Year
- Month
- Day

COLUMN 4

Date in fraction of year

COLUMN 5

Daily total sunspot number.

COLUMN 6

Daily standard deviation of the input sunspot numbers from individual stations.

COLUMN 7

Number of observations used to compute the daily value.

COLUMN 8

Definitive/provisional indicator.

- Mean: The mean value represents the average value for each attribute. The mean number of sunspots is approximately 82.95.

- Standard Deviation (std): The standard deviation quantifies the dispersion or variability in the data. The standard deviation of the number of sunspots is approximately 77.25.

- 25th Percentile (25%): The 25th percentile indicates the value below which 25% of the observations fall. The 25th percentile for the number of sunspots is 21, suggesting that a significant portion of the dataset contains lower sunspot counts.

- 50th Percentile (50% or median): The 50th percentile represents the middle value of the dataset. The median number of sunspots is 63, indicating that half of the observations have a count below this value.

- 75th Percentile (75%): The 75th percentile indicates the value below which 75% of the observations fall. The 75th percentile for the number of sunspots is 127, suggesting that a significant portion of the dataset contains higher sunspot counts.

METHODOLOGY

- <u>Data preprocessing:</u>
  - Removed NA values from the data
  - Took data only from 2010 to 2019. Based on our past experience, we have observed that the presence of cycles in the whole historical data poses a significant challenge when attempting to generate accurate forecasts.
  - Aggregated the data to have monthly frequency
- <u>Modeling:</u>
  - SARIMA
  - Exponential Smoothing
  - Return series from SARIMA model predictions.
  - ARMA model for the mean of returns.
  - GARCH model for squared residuals.
  - Forecast volatility using GARCH(1, 1)

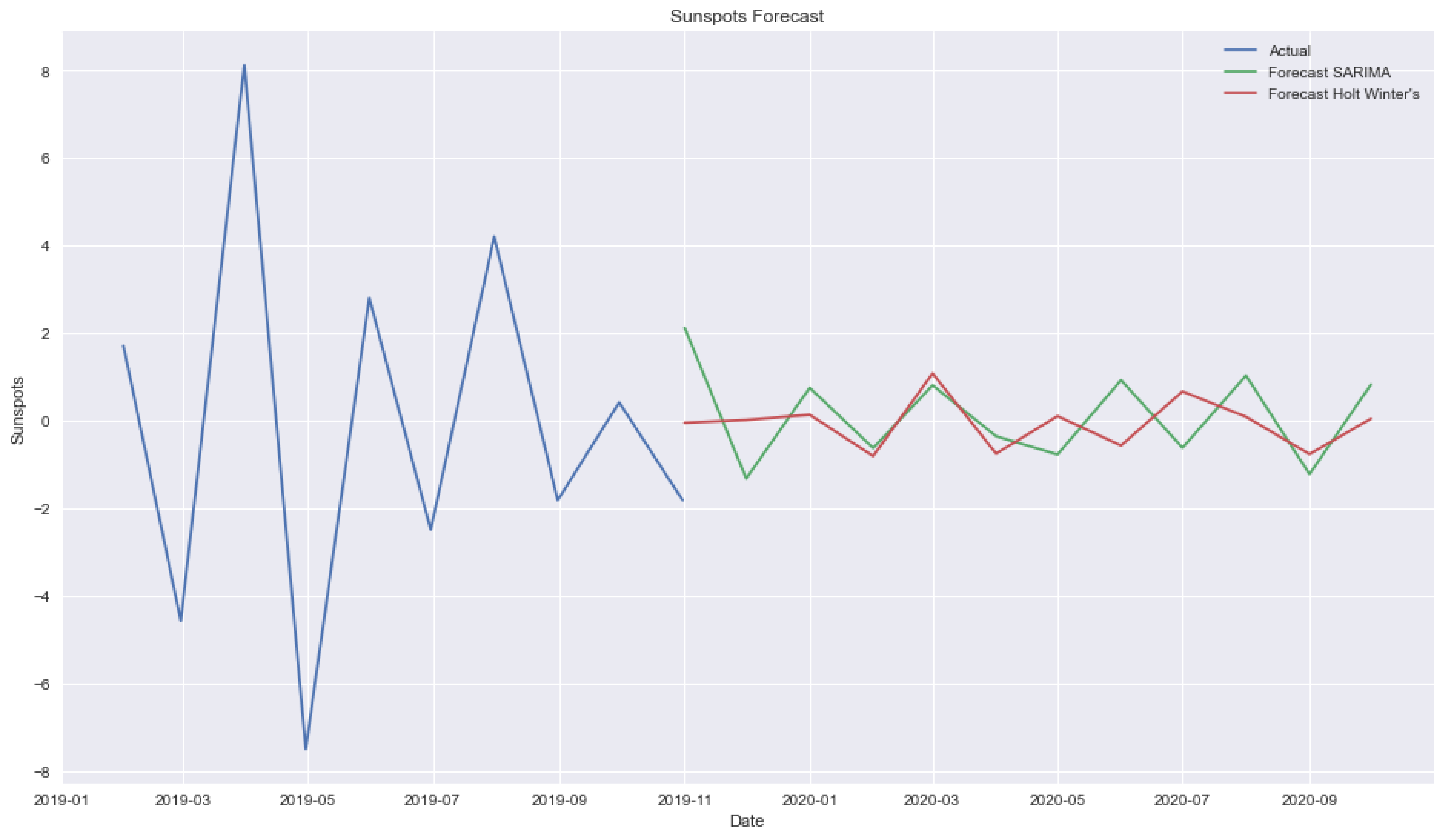
RESULTS

We have differenced the data 3 times to make it stationary.
From both manual and automated SARIMA models, we have picked the automated one as it had a lower root mean squared error of 3.79, and the manual one had 3.83.
We have modeled Holt Winter's Exponential Smoothing with the additive trend and seasonality and have got a root mean squared error of 3.76, and SARIMA has had a root mean squared error of 3.78, and as Exponential Smoothing has had a lower RMSE; then, it is a better-performing model than SARIMA.

After forecasting, we have got the following results:

RESULTS (CONT.-D)

Forecast with both SARIMA and Exponential Smoothing (1 year)



RESULTS (CONT.-D)

The analysis finds out that the series of SARIMA forecasts and its return series are both stationary, the mean dynamics of the return series are adequately captured by the ARMA(0,0) model, and the volatility dynamics are effectively modeled by the GARCH(1,1) model.

CONCLUSIONS

- With the help of SARIMA and Exponential Smoothing models, we were able to forecast sunspots for the next year.
- With the ARMA and GARCH models, we have forecasted the mean and volatility dynamics of the return series from the SARIMA prediction results of the sunspots.
- These predictions help the target audience of stakeholders enhance the research and studies of the sun's behavior and the space environment in general.