# Data Science Project: Time Series Analysis of Sunspot Data

<https://github.com/TrevorPooleData/sunspot-time-series>

## Research Question

How can historical sunspot data from 1700 to 2024 be explored and visualised to assess data quality, and how can a model be developed to forecast sunspot activity from 2025 to 2050?

## Executive summary

This project addresses the challenge of forecasting long-term solar activity using historical sunspot data spanning 1700 to 2024. Sunspots, which follow an approximately 11-year cycle, are key indicators of solar variability and have implications for climate systems, satellite operations, and power infrastructure. The dataset, sourced from the SILSO network at the Royal Observatory of Belgium, includes both reconstructed and observed annual sunspot numbers.

The approach involved rigorous data preparation, including handling missing values and assessing data quality. Exploratory analysis revealed cyclical patterns and periods of reduced activity, such as the Dalton Minimum. Confidence intervals were calculated to assess statistical uncertainty, and time-series diagnostics confirmed the data’s suitability for forecasting.

The forecasting approach was simplified by using auto-SARIMA, which automatically selected optimal model parameters based on information criteria. This removed the need for manual configuration while maintaining strong predictive performance. Benchmarking against mean-based forecasts confirmed the model’s suitability for long-term sunspot trend analysis. Visualisations were used to support interpretation and communicate uncertainty.

Key outcomes include a validated forecasting model for sunspot activity from 2025 to 2050, insights into historical solar cycles, and recommendations for future work using finer-grained data and interactive dashboards. The project showcases how historical data and statistical modelling can be combined to understand and anticipate solar variability.

## Introduction/project background

Understanding long-term solar activity is essential for anticipating its effects on climate systems, satellite operations, and power infrastructure. Sunspots—dark, magnetically active regions on the Sun’s surface—are a primary indicator of solar variability and, as noted by Hathaway (2015), follow an approximately 11-year cycle. This project aims to forecast annual sunspot numbers from 2025 to 2050 using historical data spanning 1700–2024, sourced from the Royal Observatory of Belgium.

The motivation for this project stems from personal curiosity about the historical record of sunspot observations and the potential to forecast future sunspot numbers. It also offers an opportunity to explore time series forecasting methods and assess how well they can model long-term solar activity. Investigating how this data has been collected over centuries—and whether meaningful predictions can be made from it—provides both an engaging analytical challenge and a chance to contribute to broader understanding of solar variability. The historical context of sunspot observations, dating back to 1700, provides a rich foundation for long-term analysis. By combining rigorous data preparation with interpretable modelling, this work contributes to the broader effort to understand solar variability and its implications for Earth systems.

## Data Source and Preparation

### Data Source

This project uses annual sunspot numbers from 1700 to 2024, sourced from the World Data Center SILSO at the Royal Observatory of Belgium. The dataset is a semicolon-separated file containing 325 rows, each representing the yearly mean total sunspot number, calculated as the arithmetic mean of daily totals for each calendar year.

The International Sunspot Number is derived using a formula introduced by Rudolf Wolf in 1851, which combines counts of individual sunspots and sunspot groups reported by a global network of observing stations. Wolf began compiling sunspot data in 1848 and later retrospectively applied his formula to earlier, inconsistently recorded observations, enabling the reconstruction of a continuous historical series. This includes reconstructed values for the early period and direct observational data from the 19th century onward (Hathaway, 2015).

In 2015, SILSO released Version 2.0 of the sunspot number series, improving accuracy and comparability across centuries and introducing statistical uncertainty measures (SILSO, 2015).

### Preparation

The file was loaded into a pandas dataframe with manually assigned column names and defined data types. The Year column was converted to datetime format and set as the index for time-series operations.

Data quality checks included identifying nulls, verifying data types, checking for duplicates, and flagging placeholder values (-1). The dataset was also reviewed for provisional entries and validate to ensure that it spanned the expected year range.

## Analysis Documentation

### Exploratory Data Analysis (EDA)

The time series of annual sunspot numbers (1700–2024) reveals a clear ~11-year solar cycle, with notable anomalies such as the Dalton Minimum (1790–1830) (Hathaway, 2015).

To assess uncertainty, 95% confidence intervals were calculated using the standard error of the mean, excluding early years (1700–1817) due to missing observation counts. These intervals were narrow relative to sunspot magnitudes, indicating high precision and robust variability. All entries were marked definitive, confirming SILSO validation.

A graph of a number of sunspot numbers

AI-generated content may be incorrect.

### Time-Series Analysis

The dataset was split into training (80%) and test (20%) sets, as shown in Figure 2.

A graph of a graph

AI-generated content may be incorrect.

An Augmented Dickey-Fuller (ADF) test confirmed non-stationarity, which was resolved through seasonal differencing with a lag of 11, as shown in Figure 3. A subsequent ADF test validated stationarity.

A graph showing a line

AI-generated content may be incorrect.

Autocorrelation and partial autocorrelation plots (Figures 3 and 4, respectively) showed no strong seasonal spikes at expected lags (11, 22, 33), suggesting irregular cycle alignment.

|  |  |
| --- | --- |
|  |  |

A mean-based naïve forecast (Figure 5) was used as a baseline, following the recommendation by Hathaway (2015) (see Appendix B).

A graph of a graph

AI-generated content may be incorrect.

Holt-Winters exponential smoothing was considered but rejected due to strong autocorrelation and irregular cycles (see Appendix B). SARIMA was selected for its ability to model both autocorrelation and seasonality.

Using the pmdarima library, an auto-SARIMA model was fitted (Figure 6) and benchmarked against the mean-based forecast. It achieved lower RMSE (59.62 vs. 63.79) and MAE (46.61 vs. 53.53), indicating better predictive performance.

A graph of a graph

AI-generated content may be incorrect.

Residual diagnostics (Figure 7) showed mostly normal distribution, though slight kurtosis and outliers were present. Autocorrelation in residuals was minimal, with minor spikes at lags 9 and 10, likely due to imperfect cycle alignment.

A group of graphs and diagrams

AI-generated content may be incorrect.

Forecasts for 2025–2050 (Figure 8) reflected expected solar cycles but featured wide confidence intervals, occasionally encompassing zero during minima. This highlights uncertainty in long-term predictions and suggests caution in interpreting low-activity forecasts.

A graph with blue lines

AI-generated content may be incorrect.

## Visualisations and Dashboards

Visualisation played a key role in communicating the findings of this project.

Figure 1 displays annual sunspot numbers from 1700 to 2024, with 95% confidence intervals included where observation counts were available. This visualisation complements the exploratory analysis by clearly illustrating the ~11-year solar cycle and anomalies such as the Dalton Minimum (1790–1830). The inclusion of confidence intervals helps convey statistical uncertainty, distinguishing robust observations from reconstructed estimates. As noted in the analysis, early years (1700–1817) were excluded from interval calculations due to missing observation counts—a decision that is visually communicated through the absence of drop bars during that period.

Autocorrelation and partial autocorrelation plots (Figures 2 and 3, respectively) were used to explore lagged relationships and seasonal patterns in the time series. While these plots did not directly inform parameter selection due to the use of auto-SARIMA, they supported model interpretation and confirmed the presence of seasonal dependencies.

Figures 5 and 6 display forecast visualisations comparing predicted sunspot numbers from two models: a mean-based naïve forecast (used as a baseline) and an auto-SARIMA model. Both models were trained on historical data and used to predict values in the test set. The resulting forecasts were overlaid with actual observed sunspot numbers, enabling intuitive visual assessment of each model’s accuracy.

Finally, the auto-SARIMA model—having demonstrated stronger predictive capability than the mean-based naïve baseline—was used to forecast sunspot numbers for the period 2025–2050, based on the entire historical dataset (Figure 8).

Python was chosen for its flexibility, reproducibility, and broad analytical capabilities, and all visualisations were created using Matplotlib and Statsmodels. Tableau and Excel were considered as alternatives. While Tableau supported the calculation of confidence intervals, it could not display time series data alongside upper and lower bounds within the same chart. Additionally, neither Tableau nor Excel could generate ACF and PACF plots or produce SARIMA forecasts.

## Discussion/Recommendations for future iterations

This project focused on annual sunspot data, which is suitable for long-term forecasting but may miss short-term fluctuations. Future work could explore monthly or daily sunspot numbers to capture finer-grained patterns and improve short-term accuracy.

While SARIMA performed well, alternative models such as Prophet could be considered. Prophet offers automated forecasting with built-in seasonality detection and may simplify model tuning while handling irregular time series.

Interactive dashboards were not developed in this iteration. Tableau lacks native SARIMA support, so a workflow would be needed to process forecasts in Python and export results for visualisation. Alternatively, Python-based tools like Dash or Streamlit could support integrated, interactive graphics.

Finally, although this project uses publicly available scientific data with no personal identifiers, ethical use remains important. The early reconstructed data (1700–1817) may introduce bias due to observer variability and incomplete historical records. Transparency in data provenance is maintained by citing the SILSO dataset and its Creative Commons BY-NC 4.0 license. Reproducibility is supported through clear documentation of methods and model parameters, and future iterations could explore techniques to quantify uncertainty and ensure responsible communication of forecast limitations.

## References

Alkaline-ML (2023) *pmdarima: ARIMA estimators for Python*. Version 2.0.4. PyPI. Available at: <https://pypi.org/project/pmdarima/> (Accessed: 23 August 2025).

Clette, F. & Lefèvre, L. (2015). SILSO Sunspot Number V2.0. Published by WDC SILSO - Royal Observatory of Belgium. [astro.oma.be/doi/ROB-SIDC-SILSO\_SunspotNumberV2.html](https://www.astro.oma.be/doi/ROB-SIDC-SILSO_SunspotNumberV2.html)

Hathaway, D. H. (2015). *The Solar Cycle*. Living Reviews in Solar Physics, 12(1), 4. <https://doi.org/10.1007/lrsp-2015-4>

Selva Prabhakaran (n.d.) Augmented Dickey Fuller Test (ADF Test) – Must Read Guide. Machine Learning Plus. Available at: <https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>

SILSO (2015) *Sunspot Number Version 2.0: new data and conventions*, Royal Observatory of Belgium. Available at: <https://www.sidc.be/SILSO/newdataset>

## Appendix A – Source Data

**Yearly Mean Total Sunspot Number**

Time Range: 1700–2024 (last elapsed year)

Data Description

The yearly mean total sunspot number is calculated as the arithmetic mean of daily total sunspot numbers across all days in a given year.

Note: In early years—particularly before 1749—the yearly means are based on only a fraction of the days due to missing observations.

* A value of -1 indicates missing data (i.e., no observation available).

Error Values

The yearly standard deviation is derived from daily values using the same formula applied to monthly means:

  σ(m) = √[ Σ N(d) × σ(d)² / Σ N(d) ]

Where:

* σ(d) is the standard deviation for a single day
* N(d) is the number of observations for that day

Standard Error on Yearly Mean

The standard error of the yearly mean is computed as:

  Standard Error = σ / √N

Where:

* σ is the yearly standard deviation
* N is the total number of daily observations in the year

Note: This standard error reflects precision—how sensitive the yearly mean is to random errors in daily values.

The absolute accuracy (uncertainty on the mean) is assessed over longer time scales and is not provided for individual years.

Data file structure

Filename: SN\_y\_tot\_V2.0.csv

Format: Comma Separated values (adapted for import in spreadsheets)

The separator is the semicolon ';'.

Contents:

Column 1: Gregorian calendar year (mid-year date)

Column 2: Yearly mean total sunspot number.

Column 3: Yearly mean standard deviation of the input sunspot numbers from individual stations.

Column 4: Number of observations used to compute the yearly mean total sunspot number.

Column 5: Definitive/provisional marker. '1' indicates that the value is definitive. '0' indicates that the value is still provisional.

Data Source Attribution

Sunspot data used in this analysis is provided by the World Data Center SILSO, Royal Observatory of Belgium, Brussels.

DOI: [10.4414/qnza-ac80  
Website: <https://www.sidc.be/SILSO/infosnytot>  
License: Creative Commons BY-NC 4.0 (Non-commercial use with attribution)

**Why Sunspot Numbers Exist but the Number of Observations Is Missing (1700–1817)**

The SILSO dataset includes annual sunspot numbers from the year 1700 onwards. However, for the period 1700 to 1817, the column indicating the number of observations is marked as -1, which signifies missing or unavailable data. This discrepancy arises from the nature of historical record-keeping and the methods used to reconstruct early solar activity.

Sunspot Numbers: Reconstructed from Historical Records

Although systematic solar observations began in the early 17th century, they were conducted by individual astronomers using varied instruments and methods. These records include:

* Drawings and logs by observers such as Staudacher and Schwabe.
* Written descriptions in scientific publications and personal correspondence.
* Compilations by researchers like Hoyt and Schatten, who developed the Group Sunspot Number series.

These historical sources allowed researchers to estimate sunspot numbers retrospectively. The SILSO team later calibrated and integrated these reconstructions into the modern International Sunspot Number Version 2.0, released in 2015.

Missing Observation Counts: Lack of Standardised Metadata

During the 18th and early 19th centuries, there was no centralised or standardised system for recording solar observations. As a result:

* The exact number of daily or monthly observations contributing to each annual sunspot number is often unknown.
* Many original records were incomplete, informal, or have been lost over time.
* The SILSO dataset uses -1 as a placeholder to indicate that this metadata is missing.

This absence does not affect the availability of sunspot numbers themselves, which are derived from the best available historical evidence.

Reference:

Clette, F., Lefèvre, L., Cagnotti, M., Cortesi, S. and Bulling, A., 2016. *The revised sunspot number: assembling all corrections*. Solar Physics, 291(9-10), pp.2629–2651.

## Appendix B – Time-Series Analysis

### Stationarity and the Augmented Dickey-Fuller (ADF) Test

What is Stationarity?

Stationarity is a fundamental concept in time series analysis. A stationary time series has statistical properties—such as mean, variance, and autocorrelation—that do not change over time. Ensuring stationarity is essential for reliable forecasting using models such as ARIMA or SARIMA.

Key indicators of non-stationarity include:

* Trend: A long-term upward or downward movement in the data.
* Seasonality: A repeating pattern at regular intervals (e.g., annually, monthly).
* Changing autocorrelation: Relationships between observations that vary over time.

The Augmented Dickey-Fuller (ADF) Test

The ADF test is a statistical method used to determine whether a time series is stationary by testing for the presence of a unit root.

Hypotheses:

* Null hypothesis (H₀): The time series has a unit root → non-stationary
* Alternative hypothesis (H₁): The time series does not have a unit root → stationary

Interpreting the p-value:

* p-value < 0.05: Reject H₀ → The series is stationary
* p-value ≥ 0.05: Fail to reject H₀ → The series is not stationary

This test is particularly useful for natural phenomena such as sunspot cycles, which often exhibit long-term trends or periodic behaviour.

Reference:  
Selva Prabhakaran (n.d.) *Augmented Dickey Fuller Test (ADF Test) – Must Read Guide*. Machine Learning Plus. Available at: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/ (Accessed: 23 August 2025).

### The Role and Limitations of the Mean-Based Naïve Forecast

The mean-based naïve forecast assumes that future sunspot numbers will remain consistent with the historical average. While this method is not suitable for forecasting cyclical solar activity, it serves as a useful **benchmark** or **null model** against which more sophisticated approaches can be evaluated.

Its simplicity allows it to act as a baseline: if a forecasting model cannot outperform the mean-based forecast, it offers limited practical value. However, due to the following limitations, it is not appropriate for modelling sunspot behaviour:

* **Cyclical Nature Ignored**: Sunspot activity follows a well-established ~11-year cycle. The mean forecast disregards this periodicity, leading to inaccurate predictions.
* **Failure to Capture Seasonality**: Although the ADF test confirms that the data is stationary, the mean-based approach does not account for the strong seasonal and cyclical components inherent in solar activity.
* **Lack of Predictive Skill**: It cannot model the peaks, troughs, or transitions between solar cycles, which are critical for understanding and forecasting sunspot behavior.

As Hathaway (2015, §7.2) notes, the mean amplitude serves as a **benchmark for prediction methods** but represents a **"prediction without any skill."** If other models cannot outperform this baseline, they offer little practical value.

**Conclusion**: The mean-based naïve forecast is included solely as a reference point. Its poor performance highlights the value of more sophisticated models like **SARIMA**, which incorporate seasonality and autocorrelation to better reflect the underlying structure of the data.

**Reference**: Hathaway, D. H. (2015). The Solar Cycle. Living Reviews in Solar Physics, 12(1), 4. <https://doi.org/10.1007/lrsp-2015-4>

### Why Holt-Winters Exponential Smoothing Is Not Suitable for Sunspot Data

Limitations of Holt-Winters

1. Strong Autocorrelation  
   The raw sunspot data exhibits significant autocorrelation across multiple lags, particularly around the 11-year cycle. Holt-Winters assumes residuals are uncorrelated, making it unsuitable for data with long memory or autocorrelated errors. According to Hyndman & Athanasopoulos (2021), exponential smoothing methods do not attempt to model the autocorrelations in the data. They are best suited to short-term forecasting where autocorrelation is not a major concern.
2. Irregular Cycles  
   Sunspot activity follows a quasi-periodic cycle — not strictly additive or multiplicative. Holt-Winters performs best with stable, regular seasonal patterns, which this data does not exhibit.
3. Underfitting Risk  
   Holt-Winters lacks autoregressive and moving average components, which can lead to underfitting. This results in poor forecasts and residuals that still contain structure.
4. SARIMA is More Appropriate  
   SARIMA explicitly models autocorrelation and seasonality, making it better suited for time series such as sunspots that exhibit cyclical behaviour and correlated errors.

Conclusion

Holt-Winters is not recommended for forecasting sunspot data due to its inability to handle autocorrelation and irregular cycles. SARIMA provides a more robust framework for capturing the underlying structure of the data.

Reference

Hyndman, R.J. and Athanasopoulos, G. (2021) *Forecasting: Principles and Practice*. 3rd edn. OTexts. Available at: <https://www.statlearning.com/>  (Accessed: 23 August 2025).

### Seasonal Autoregressive Integrated Moving Average (SARIMA)

Introduction to ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a widely used model for forecasting time series data. It combines:

* Autoregression (AR) – Uses past values to predict future ones.
* Integration (I) – Applies differencing to make the series stationary.
* Moving Average (MA) – Models the relationship between an observation and past forecast errors.

ARIMA is suitable for data with trends but not for data with seasonal patterns.

Why SARIMA?

SARIMA extends ARIMA by incorporating seasonal components, making it suitable for time series data that exhibit regular cycles. It models future values based on:

* Differenced observations
* Error terms
* Seasonal differences

In this project, SARIMA was selected to account for the ~11-year solar cycle observed in the sunspot data.

Advantages of SARIMA

* Captures both short-term and seasonal patterns in time series data.
* Explicitly models autocorrelation and seasonality

SARIMA Parameters

SARIMA requires two sets of parameters:

* Non-seasonal order: *(p, d, q)*
  + *p*: Number of autoregressive terms
  + *d*: Number of non-seasonal differences
  + *q*: Number of moving average terms
* Seasonal order: *(P, D, Q, M)*
  + *P*: Seasonal autoregressive order
  + *D*: Number of seasonal differences
  + *Q*: Seasonal moving average order
  + *M*: Number of observations per season (e.g. 11 for sunspot data with an 11-year cycle)

These parameters enable SARIMA to model both short-term dynamics and seasonal cycles effectively.

Auto SARIMA Model Summary

A screenshot of a computer program

AI-generated content may be incorrect.