Predicting Sunspots using Machine Learning

By ELizabeth Tyree

Problem Statement and Outline

Is the frequency of sunspots predictive of the frequency of earthquakes?

Importance: Why Sunspots and Earthquakes?

Data: Solar Cycle Characteristics and Feature Importance

Approach: Machine Learning Methods ARMA, SARIMAX, and LSTM

Problem Statement and Outline

Is the frequency of sunspots predictive of the frequency of earthquakes?

Results: The Lessons Learned

Evaluation and Performance

Conclusions and Recommendations

Next Steps

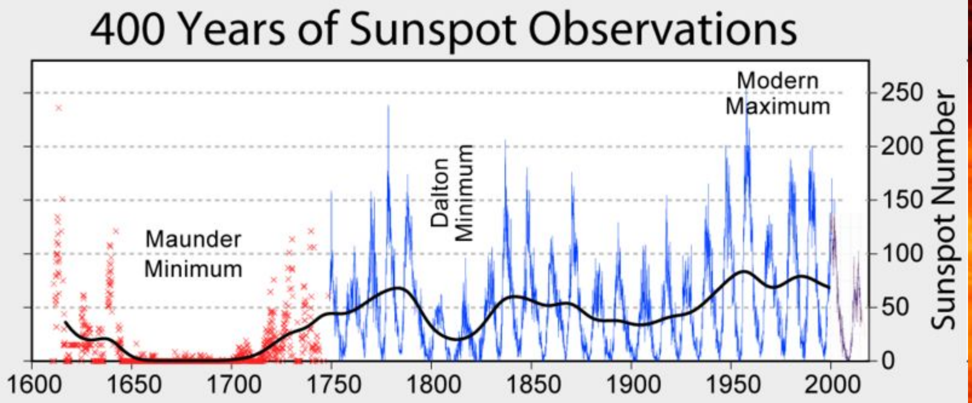
Importance: The ability to better understand Earthquakes, can help save lives.

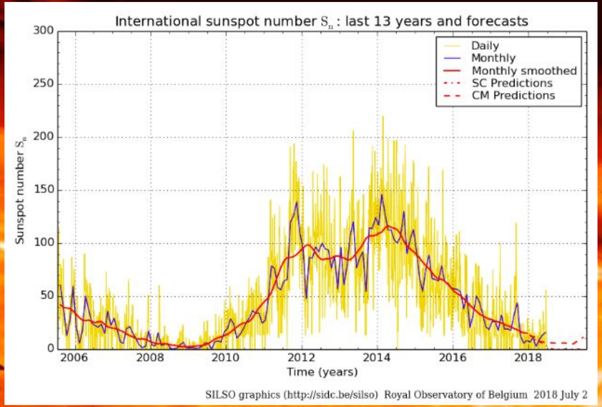
(Credit: Stephen M. Wessells, U.S. Geological Survey. Public domain.)

Data: Solar Cycle Characteristics

Sunspot data source: SILSO data, Royal Observatory of Belgium, Brussels.

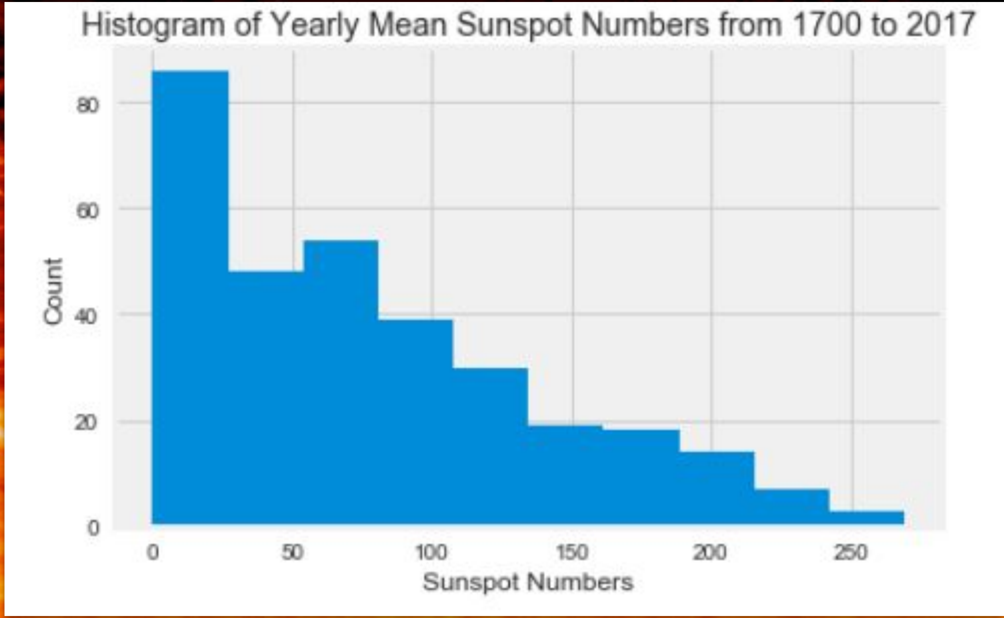
- Sinusoidal in nature, not at all linear.
- Solar cycles include:
    - 27 day rotation of the sun
    - 1 Cycle is 11 years (average) = 4015 days.
    - 89 year Cycle = 32485 Days.
- Solar data is noisy. In this context anything that obscures or detracts from the solar cycle pattern is considered noise.



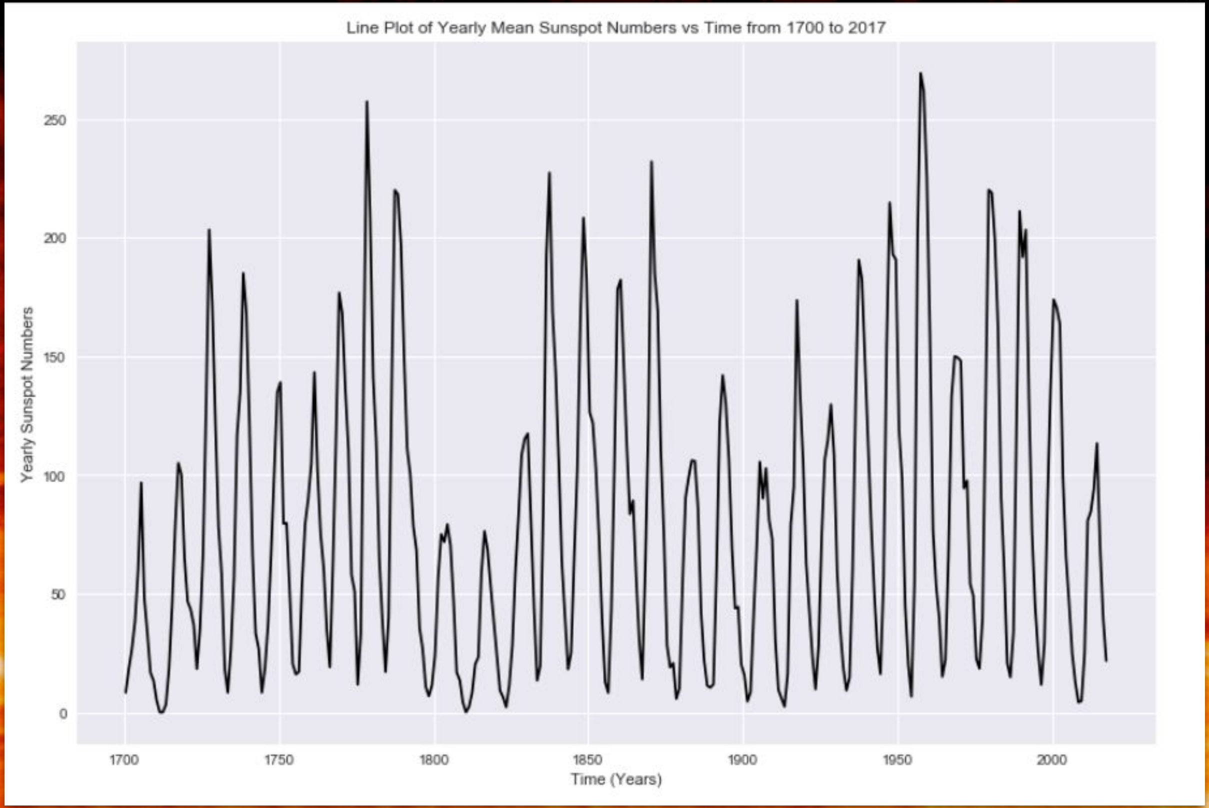


The Distribution of Sunspots Contained in the Data

Sunspot data source: SILSO data, Royal Observatory of Belgium, Brussels.



2 Features: Yearly Mean Sunspots, and Time



The Approach:

- Data: Yearly Mean Sunspots from 1700 to 2017.
  - Removed data flags from sunspots.
  - Decomposed the Timestamp (looked at Seasonality and Trends in the data)
- Tools
  - In order to investigate non-linear events over time, time series based models were used. These models are the ARMA, SARIMAX, and LSTM.
- What the Models found (aka model evaluation).

About The Models:

1. The Autoregressive Moving Average (ARMA)

2. Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model (SARIMAX)
3. Recurrent neural network
    a. Long-short term memory (LSTM)

Results: Metrics Used

Mean Absolute Error (MAE)

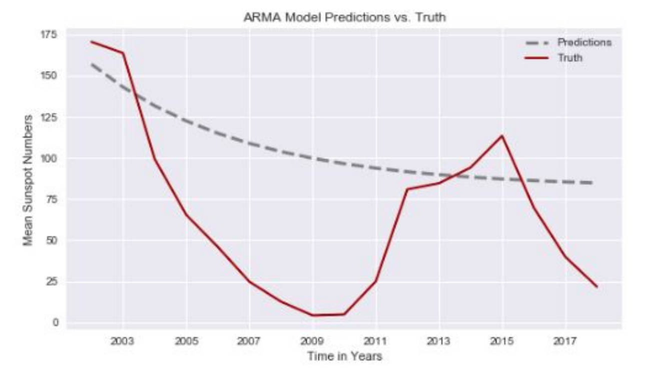
- The total error between my prediction and my test data.
- The lower the number the better!

Root Mean Square Error (RMSE)

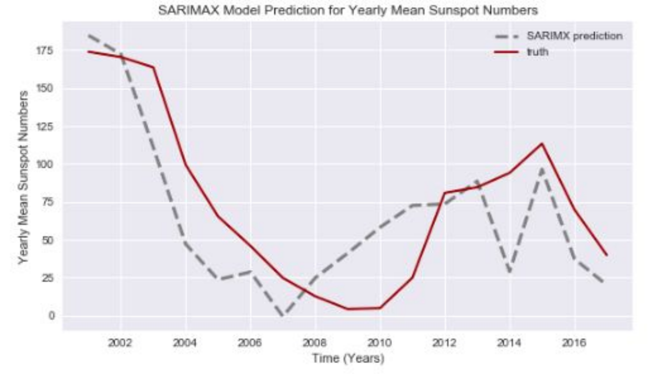
- The lower the number the better!

Model Results: Mean Absolute Error and Root Mean Square Error.

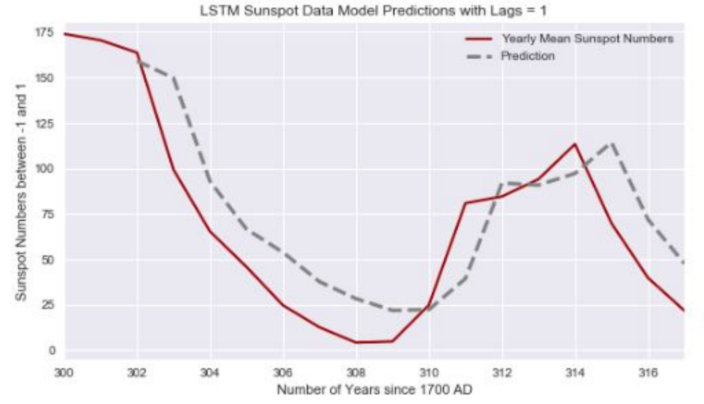
ARMA
MAE = 46.8
RMSE = 56.6



SARIMAX
MAE = 18.9
RMSE = 26.9



LSTM
MAE = 51.1
RMSE = 64.0



Evaluation and Performance

- The SARIMAX numerically performed the best, had the lowest scores of all three Models. (This was a surprise!)
- Both the SARIMAX and LSTM were good at capturing the overall solar cycle, but the LSTM was the best at capturing some of the finer details of the high's and lows of the solar data out of all 3 Models.
- The ARMA performed the worst but had better scores than the LSTM.

Evaluation and Performance

So what does this all mean?

- ARMA is a bad Model for Sunspots.
- The LSTM and SARIMAX performed the best at capturing the solar cycle.
- SARIMAX or LSTM, which is better?

Conclusions and Recommendations

Most important:

- Understanding time series based tools when applied to noisy, high variability data with hard to predict seasonality and trends.

Recommend redoing analysis with Monthly and Daily solar data.

- Add other performance metrics.
- Signal Analysis as a performance metric.

Next Steps:

Is the frequency of sunspots predictive of the frequency of earthquakes?

Phase 1: Sunspot Modeling using ARMA, SARIMAX, and LSTM

Phase 2: Earthquake Modeling using SARIMAX, and LSTM

Additional Slide

```
MAE =  51.111953770287016
RMSE =  64.04173279482075
```