Predicting the Number of Sunspots Per Month and Per Quarter Using ARIMA Models

Suvrath Arvind, Clayton Greenberg

Abstract

The number of sunspots in a given year changes as the sun goes through solar cycles, with peaks happening at regular intervals. When these peaks are plotted, a curve appears, similar to the oscillating sinusoidal wave. Because of its oscillatory nature, predictions of future sunspot values could be found since it is safe to assume that the number of sunspots would always follow a pattern. However, a simple, ordinary sine function, or any algebraic function for that matter, would not allow us to plot and predict future data points due to the complexity of the curve at hand. This led us to the hypothesis that in order to predict the future number of sunspots, models that involve autoregressive and moving average components (namely the ARIMA model) would be the most effective. In order to measure effectiveness, the mean-squared error (MSE) would be used, with a lower value (closer to 0) meaning better fit. The reason why we chose these sophisticated models was because these models took into account prior data points and their trends and seasonality to predict future data points. This essentially meant that this model would predict based on prior points, not on a fixed point or equation, like the sine curve. After plotting all of these models and finding the MSE for each, we drew the conclusion that the ARIMA model proved to produce the most accurate curve, with a MSE of only 460, as compared to the MSE that the best sine curve could produce: 21 million.

1 Introduction

Although there is a pattern in the number of sunspots, there is still uncertainty with this data since the peaks of this data follow their own oscillation pattern, whereas the troughs of this data follow a completely different oscillation pattern. This uncertainty prevents us from using any simple curve to fit the data accurately and requires us to use more complicated models involving autoregression and moving averages. However, even these models have levels of sophistication and accuracy and with models using different components, all models must be tried to see which one gives us the best result. The most basic of these models are the AR and the MA models, where “AR” stands for autoregression and “MA” stands for moving average, which only use parts of the complete ARIMA model.

Sunspot data is like any natural process, it is hard to predict using simple tools. Data collected over hundreds of years is necessary to make any predictions due to the robust requirements of this data. Additionally, this data does not follow a single pattern, but instead is a combination of multiple patterns, each meaning something different. For instance, in the data, the amount of time between any two peaks is 11 years, on average, but the amount of time for two peaks in the data to be equal takes about 43 years. This limits us to possible curves to use and nearly requires us to use sophisticated models to predict the data and fit a curve that is similar to that produced by initial data points.

Lastly, the sunspots had to be associated with a specific time interval. Based on the nature of the data provided, it was convenient to use monthly data as well as quarterly data to plot the curves. Although this data would look slightly different than annual data, it would provide

greater accuracy to the curve by plotting data points that were associated with a smaller time interval instead of larger time intervals which would limit the number of data points on the graph. The only real drawback of this method was the look of the graph since using smaller subdivisions of time on the graphs would make the graph have a rougher shape than when annual data was plotted.

2 Background

Attempts have been made to produce models regarding future values of the number of sunspots. For instance, the NOAA's Space Weather Prediction Center has produced models that use data as recently as September 2022 to predict the number of sunspots per month[1]. This organization created a curve that used prior data points to predict future values, similar to that of this project. The goal of this organization was to not only predict the number of sunspots, but also use this data to determine the lifetimes of low orbiting satellites. This is a key application of this prediction and is essential for the growth of the modern world, one that requires satellites for everyday functions.

3 Dataset

The data that was used was produced by Inspirit AI[2] and categorized the number of sunspots recorded on specific days. The data converted this day into a decimal representation (accuracy up to a thousandth of a year) in order for the data to be usable. However, this produced 70471 data points, which meant training on this many data points would be too time consuming and could potentially crash the computer and program, rendering the model to be useless. In order to fix this, these decimal representations (labeled "Date in Fraction of a Year") were rounded to the nearest year, nearest month (which was taken to be a tenth of a year) and nearest quarter in order to produce a model that would work effectively. Even though this would not be as accurate, it would be accurate enough to justify if a model was a good fit or not. This rounding shrank the data set to have under 5000 data points, something that would work well for an ARIMA model to train on. Below are the tables representing the original data and the table representing the quarterly data.

[1] https://www.swpc.noaa.gov/products/solar-cycle-progression
[2] https://storage.googleapis.com/inspirit-ai-data-bucket-1/Data/AI%20%2B%20X/Group/Physics/SunSpot/sunspot_data.csv

Year	Month	Day	Date In Fraction Of Year	Number of Sunspots
1818	1	8	1818.021	65
1818	1	13	1818.034	37
1818	1	17	1818.045	77
1818	1	18	1818.048	98
1818	1	19	1818.051	105
...	...	...	...	...
2019	10	27	2019.821	0
2019	10	28	2019.823	0
2019	10	29	2019.826	0
2019	10	30	2019.829	0
2019	10	31	2019.832	0

Fig 1: Original data; it has over 70 thousand rows, so this data is not usable

	Rounded_quarterly	Number of Sunspots
0	1818.00	961
1	1818.25	2586
2	1818.50	4536
3	1818.75	2513
4	1819.00	1493
...	...	...
803	2018.75	531
804	2019.00	427
805	2019.25	863
806	2019.50	104
807	2019.75	46

808 rows × 2 columns

Fig 2: Quarterly data; it has only 808 rows, so this data is much better to use

As the two tables show, the compression of this data set was a key step in the building of this model. Without this step, training a model would take a very long time and would not get us anywhere since the dates during which sunspots were recorded were not regular. Rounding the data to the nearest tenth or quarter allowed there to be some sort of consistency in the data. Although the data was rounded to the nearest year, this would lead to too few points for the ARIMA model to train on and would not accurately represent the data. This rounding step allowed for different models, such as the ARIMA model, AR model, or the MA model (all of which will be discussed in the next section), to train on the data in order to create a prediction.

4 Methodology / Models

4.1 AR/MA/ARIMA Models

ARIMA models and other autoregressive or moving average models have been used to model data like that of the number of sunspots. Because of the uncertainty of the data, models like ARIMA, and even AR, MA, or ARMA (the components that make up the ARIMA model), can

be used to predict sunspot data as these models take into account prior data points to make predictions about future ones.

Autoregressive models, commonly referred to as AR models, use regression to predict future points. These models use prior data points to predict future ones. This model uses trends in the data, a quality that predicts long term behavior of a graph (such as increasing or decreasing).[3] Although this quality makes this model powerful, we believe that it limits what this model can do. This model does not take into account any errors produced between the model and the original data it trains on and can lead to unfavorable results in the prediction.

Moving Average models, commonly referred to as MA models, use residuals between the prediction and the actual data points in order to predict future data points. This means that the model tries to "correct" itself each time, but it does not make predictions based on previous values. This is the key difference between AR and MA models: one uses past values to predict future ones while the other uses past errors to produce values that would try to have less error. Additionally, unlike AR models, MA models are used to see the seasonality of the data. Seasonality is similar to a trend, but unlike a trend, it describes data points over short intervals, analogous to a period of an oscillating object. It would be more appropriate to calculate the seasonality of the sunspot data as the data points oscillate up and down with the maxima forming patterns. The MA model would accurately describe how the data varies over short intervals and can be used to make conclusions about the oscillatory nature of the data.[4] [5]

Although MA models are effective when analyzing this data, using an ARIMA or SARIMA (where the "S" stands for Seasonal and takes into account specific seasonal parameters) model would allow for complete analysis of the data. These models would take into account both trends and seasonality and can predict where future data points will lie. These models are significantly more sophisticated and, in some cases, it can even crash the computer when running. However, depending on the data and the parameters put into the model, it can work quickly and provide an accurate model.

In order to measure the success of a model, three factors would be used to determine the accuracy: the MSE (Mean Squared Error), the AIC (Akaike Information Criterion), and the Logarithmic Likelihood. These values are essential for the determination of the success of each model. For a model to be a good fit, the MSE should be low, the AIC should be low and the Log Likelihood should be high. When testing the AR and MA models, the MSE was not even checked because of how unfavorable the AIC and Log Likelihoods were for these models. However, since the AR and MA models were not that strong and produced, the MSE was not calculated due to the unfavorable results produced by the AIC and the Log Likelihood. This decision was made since these two types of models resulted in a high AIC and a low Log Likelihood.

4.2 Model From Scratch

Another approach that was taken to solve this problem was to create models from scratch, using simple sine functions. These models took into account the following parameters:

- Amplitude
- Period

[3] https://towardsdatascience.com/time-series-models-d9266f8ac7b0
[4] https://365datascience.com/tutorials/time-series-analysis-tutorials/moving-average-model/
[5] https://towardsdatascience.com/defining-the-moving-average-model-for-time-series-forecasting-in-python-626781db2502

- Phase Shift
- Vertical Shift

Two models were created, one that involved one sine function and another that multiplied two sine functions together; in order to determine the accuracy of each model, the MSE would be compared (since other measurements could not be done on a model created from scratch).

In the first model, that only had one sine function, was an altered form of the simple sine function, $f(x) = sin(x)$, and required four parameters:

- Amplitude, represented by A
- Horizontal Compression (which resulted in a change in the period), represented by p
- Phase Shift (which translated the function horizontally), represented by s
- Vertical Shift, represented by v

With these parameters, the sine function would resemble: $-Asin(\frac{2\pi}{p}(x - s)) + v$.

Similar to the first model, the second model's basis was the sine function, but this model would use two sine functions being multiplied together. Additionally, the absolute value of the first sine function that was being multiplied was taken in order to reduce variability in the data. Unlike the first model, this model had more unknowns and took into account six parameters:

- Amplitude, represented by A
- Vertical Shift, represented by v
- Period for First Sine Function, represented by p_1
- Period for Second Sine Function, represented by p_2
- Phase Shift for First Sine Function, represented by s_1
- Phase Shift for Second Sine Function, represented by s_2

This made the function resemble: $-A\left|(sin(\frac{2\pi}{p_1}(x - s_1))\right| \times (sin(\frac{2\pi}{p_2}(x - s_2)) + v)$

In order to find the best parameters, nested for-loops would be used changing all of the factors in order to create as many combinations as needed and the MSE of each iteration would be compared against all other trials. The lowest MSE would be printed out along with the best parameters, giving us the best sine curve.

5 Results

5.1 Results from Sine Function Models

The two models created using the sine functions were compared by checking which one had the lowest MSE. The first model (the one that used only one sine function) proved to have the best MSE when it resembled the function y = $-5000sin(\frac{2\pi}{11}(x - 1818)) + 7275$. Below is the graph that represents this curve, alongside the original data:

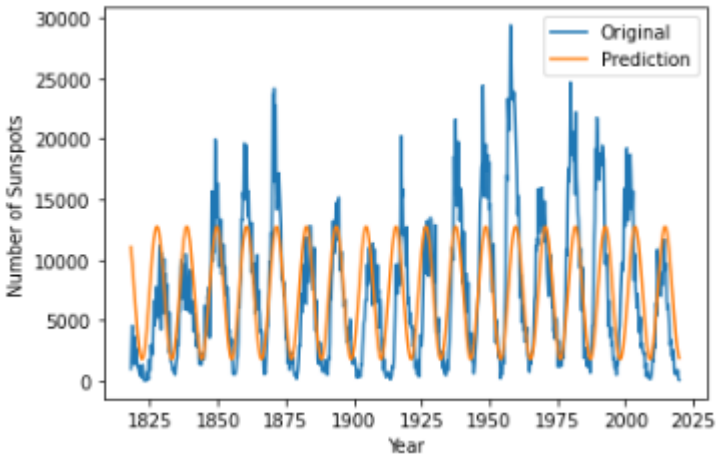

Fig 3: First Sine Model plotted with Original Data

Even though this clearly does not fit the data well, it was the best graph produced by the single sine function. This prediction resulted in the MSE being 20,504,187.151.

The second model proved to have the best MSE when it resembled the function $y = -5000\left|(sin(\frac{2\pi}{43}(x - 1818))\right| \times (sin(\frac{2\pi}{11}(x - 1818)) + 7275))$. This MSE was 22,002,786.174 and proved that this model did poorly as compared to the first model, the one that only had a single sine function.

The results of the two sine functions showed that the first model predicted the data better than the second model, but even then, the predictions were far off, with the MSE being very high. This meant that there had to be a better model to predict the data, leading to the ARIMA model being used to represent the data.

5.2 Results from ARIMA Model

The ARIMA model proved to be a much better option for the prediction. This model is much more powerful than the sine function model, but it uses more memory to run. Below is the graph of the ARIMA model compared to the original data.

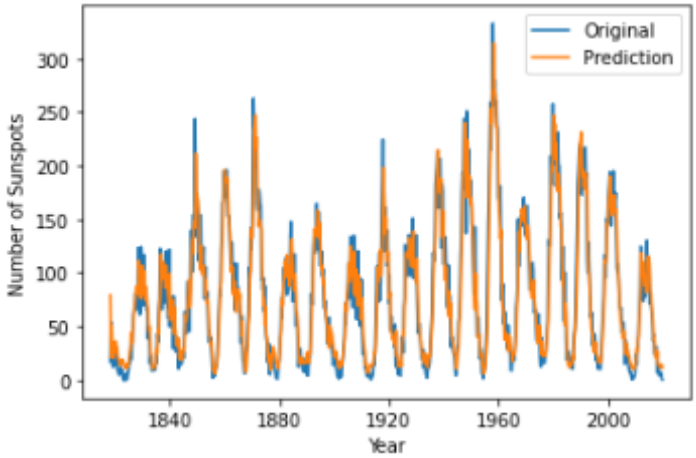

Fig 4: ARIMA Model plotted alongside the original data

As this graph shows, the ARIMA model came much closer to the original data as compared to the previous models. It produced a Mean Squared Error of only 460.245! This meant that this model was much more accurate in predicting data points than a sine function because this model based predictions off of previous data points, not based on constants. Below, is the table summarizing the results of this model:

SARIMAX Results

Dep. Variable:	Number of Sunspots	No. Observations:	808
Model:	ARIMA(3, 0, 3)	Log Likelihood	-3621.130
Date:	Sun, 09 Oct 2022	AIC	7258.261
Time:	17:42:22	BIC	7295.817
Sample:	03-31-1818	HQIC	7272.682
	- 12-31-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	79.1797	6.003	13.191	0.000	67.415	90.945
ar.L1	1.5920	0.153	10.396	0.000	1.292	1.892
ar.L2	-0.3037	0.298	-1.019	0.308	-0.888	0.281
ar.L3	-0.3170	0.149	-2.133	0.033	-0.608	-0.026
ma.L1	-0.9402	0.150	-6.257	0.000	-1.235	-0.646
ma.L2	-0.1882	0.199	-0.947	0.344	-0.578	0.201
ma.L3	0.3099	0.068	4.577	0.000	0.177	0.443
sigma2	455.5388	18.344	24.833	0.000	419.585	491.492

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	109.92
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.94	Skew:	0.50
Prob(H) (two-sided):	0.63	Kurtosis:	4.51

This table shows that the Log Likelihood was a relatively low negative number (low in magnitude) and the AIC was a large positive number. This meant that this model proved to work really well in predicting the results for the sunspot data, which, as the MSE of the sine function showed, was as simple as a single sine function.

Lastly, a SARIMAX model (a model that uses SARIMA) was used. Unlike the ARIMA model, this model took into account the seasonality of the data and was even more powerful than the ARIMA model. However, this used up a large amount of memory, even resulting in the program crashing at times, and took a large amount of time to run. Its functionality was similar to that of the ARIMA model as it still used autoregressive and moving average models to predict data points. Below is the graph of this model compared to that of the original data set.

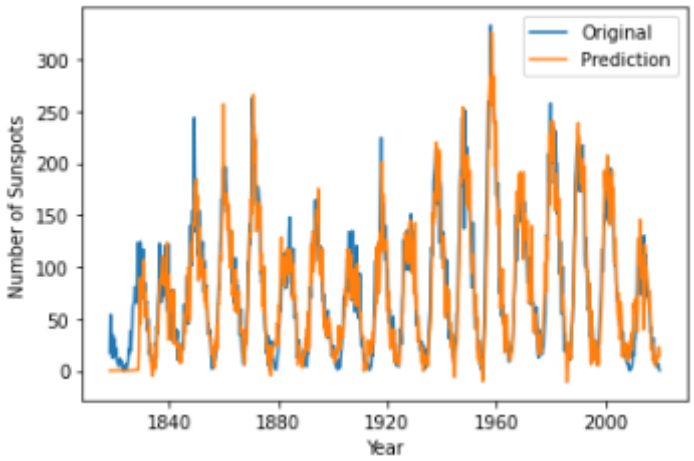

Fig 5: SARIMAX Model plotted alongside original data

As this graph shows, this model also came close to the original data set, but it was not as accurate as the ARIMA model. This could be seen in the MSE of this model, which was about 719. This meant that a more powerful model did not necessarily mean a more accurate model and that the ARIMA model was the best option. The ARIMA model took only seconds to run, whereas the SARIMAX model took nearly five minutes to run. This means that not only did the ARIMA model provide a better prediction, it also took a significantly shorter amount of time to run. However, when comparing the other measurements of accuracy (like AIC and Log Likelihood), this model proved to work as well as the ARIMA model. Below is the table that summarizes this model:

SARIMAX Results

Dep. Variable:	Number of Sunspots	No. Observations:	808
Model:	SARIMAX(3, 0, 10)x(1, 1, [], 43)	Log Likelihood	-3570.579
Date:	Sun, 09 Oct 2022	AIC	7171.158
Time:	17:55:42	BIC	7240.756
Sample:	03-31-1818 - 12-31-2019	HQIC	7197.950
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7668	0.174	4.401	0.000	0.425	1.108
ar.L2	0.8818	0.065	13.654	0.000	0.755	1.008
ar.L3	-0.7294	0.131	-5.563	0.000	-0.986	-0.472
ma.L1	-0.2011	0.178	-1.133	0.257	-0.549	0.147
ma.L2	-0.9249	0.136	-6.815	0.000	-1.191	-0.659
ma.L3	0.4298	0.066	6.549	0.000	0.301	0.558
ma.L4	0.2330	0.055	4.226	0.000	0.125	0.341
ma.L5	-0.0351	0.078	-0.448	0.654	-0.188	0.118
ma.L6	-0.1017	0.063	-1.624	0.104	-0.224	0.021
ma.L7	-0.0344	0.050	-0.685	0.493	-0.133	0.064
ma.L8	0.1711	0.044	3.856	0.000	0.084	0.258
ma.L9	0.0605	0.061	0.993	0.321	-0.059	0.180
ma.L10	-0.1110	0.063	-1.769	0.077	-0.234	0.012
ar.S.L43	-0.4684	0.029	-16.249	0.000	-0.525	-0.412
sigma2	664.5920	29.870	22.250	0.000	606.048	723.136

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	39.42
Prob(Q):	0.79	Prob(JB):	0.00
Heteroskedasticity (H):	1.17	Skew:	0.12
Prob(H) (two-sided):	0.21	Kurtosis:	4.09

This table shows that the AIC and Log Likelihood were similar to that of the ARIMA model and were similar to that of a strong model. However, because the MSE was much higher than the ARIMA model, this model was not as strong as the ARIMA model.

6 Conclusion

In this project, we were able to successfully create a model that could accurately predict data points in sunspot data. An ARIMA model was required to produce the best results. Predicting the number of sunspots in a given time interval would be possible only by using models such as this one and others that use autoregression and moving average models due to the high variability in the data.

In conclusion, in order to predict the number of sunspots in a given time interval, our experiments with an ARIMA model led to the lowest error. This is shown through various measures of accuracy and through graphical analysis. Unlike this model, a simple sine function would not be as accurate as any autoregressive moving average models (ARIMA or SARIMA) and the best parameters for this model resulted in an MSE fifty thousand times higher than that of the ARIMA model. Lastly, this research suggested that the additional power of a seasonal

ARIMA model does not correspond with a substantial increase in accuracy. In effect, we cannot verify that sunspots follow a 43 year seasonality pattern from this dataset.

In the future, this work could be expanded upon to predict other astronomical phenomena, including the probability of the Earth being hit by meteorites or even solar storms.

Acknowledgements

We would like to thank Polygence for allowing this project to take place. Without this, no work could have taken place and this research would not have been conducted.

References

[1] *Solar cycle progression*. Solar Cycle Progression | NOAA / NWS Space Weather Prediction Center. (n.d.). Retrieved October 15, 2022, from https://www.swpc.noaa.gov/products/solar-cycle-progression

[2] Shetty, C. (2020, September 22). *Time Series models*. Medium. Retrieved October 15, 2022, from https://towardsdatascience.com/time-series-models-d9266f8ac7b0

[3] Mehandzhiyski, V. (2021, October 20). *What is a moving average model?* 365 Data Science. Retrieved October 15, 2022, from https://365datascience.com/tutorials/time-series-analysis-tutorials/moving-average-model

[4] Peixeiro, M. (2021, December 6). *Defining the moving average model for time series forecasting in Python*. Medium. Retrieved October 15, 2022, from https://towardsdatascience.com/defining-the-moving-average-model-for-time-series-forecasting-in-python-626781db2502

[5] Srujan, K. (2020, September 1). *Predicting sunspots using Arima*. Kaggle. Retrieved October 15, 2022, from https://www.kaggle.com/code/krishnasrujan/predicting-sunspots-using-arima/notebook