Relationship Between Earth-Moon Distance and Tides

Aiden-Shangcheng Wu

Wesley-Shengwen Chang

John-Yanzhuo Bai

1. Background

Our topic is the relationship between Earth-Moon distance and tides. The reason why we choose this topic is that it is closely connected to us. Because they will cause heavy damage to human society. If we cloud predict the tides, we can prevent this damage. Taking all the advantages of tides. Tides will do more good than harm. Moon has a close relationship with tides, the distance between Erath and Moon is one of the main reasons for the formation of tides. So, we use Python to achieve data through API and CSV files at first. Luckily, both are highly predictable. That's good for us to use Python to do data prediction.

1. Project | Data Preparation

By searching the Internet, we found a website named Timeanddate, which provides detailed data on the distance between the Earth and the Moon. Therefore, through this website, we successfully obtained a total of 365 pieces of earth-moon distance data from January 1, 2022, to December 31, 2022. Then we found the tides data from the National Oceanic and Atmospheric Administration, an official United States government website. This website provides us with the height of tides.

But there are still many problems contained in these data. For the distance data, we can only achieve one data at a time and there is no accurate tutorial here that tells us how to get only distance data from their website, we found a sample code to retrieve the data, but unfortunately, when we ran the sample code, it returned an empty list. So, we spent a lot of time, trying one by one, and finally determined a code that could be obtained distance. (fig.1)

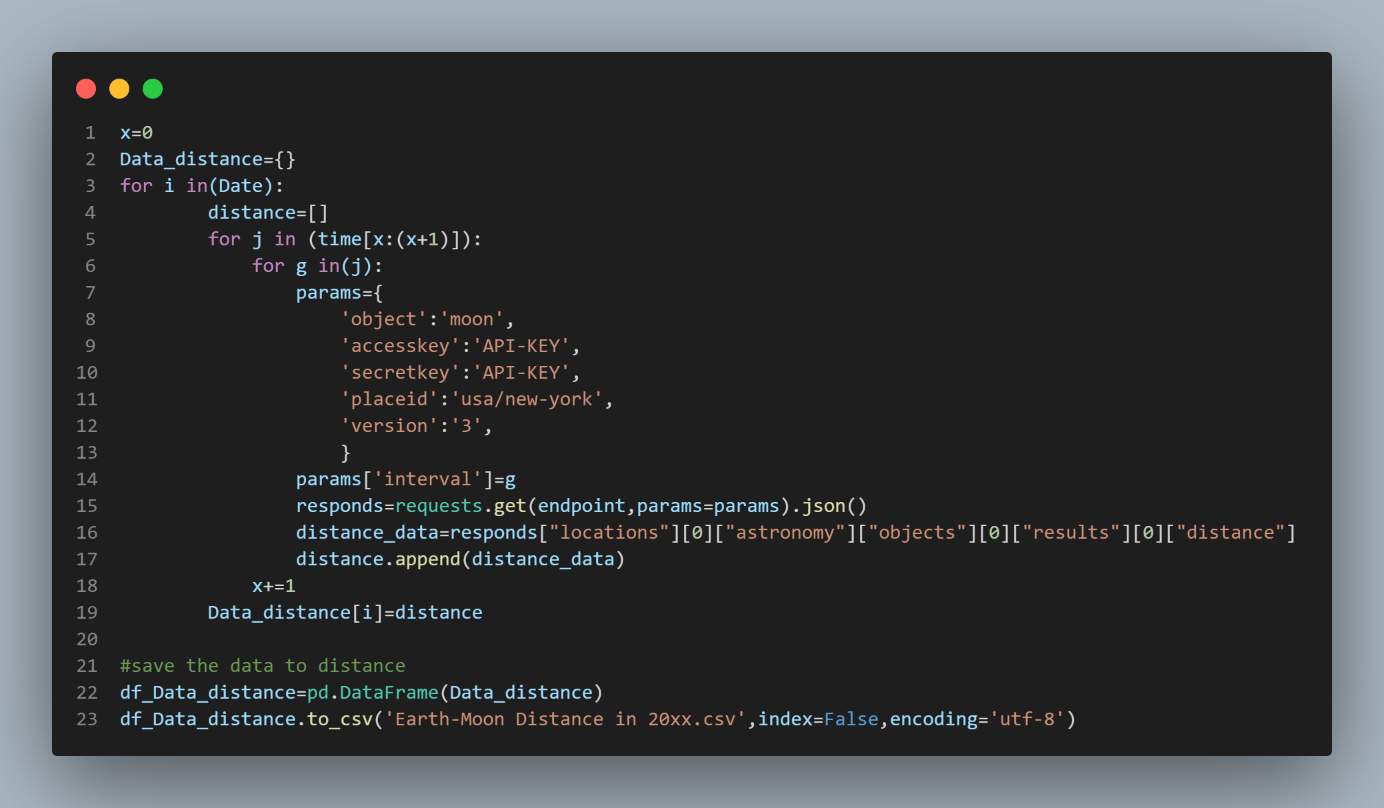


fig.1

then we use three loops in the code to make sure that the code can automatically get all the data we want and save them into one CSV file (fig.1)

For the tides data, it’s easier to get than the distance data because the website provides CSV files so that we can download them directly. However, the tides data is too accurate because it recorded the height of tides every six seconds, so, it contains 240 pieces of data a day, and 7440 pieces of data a month, which means there will be a large amount of data to be dealt with.

For data visualization, the first step is to load these raw data. Because the data are all divided into different CSV files, so after we load the data, is to come out with a way to put all the data together in a time sequence to exhibit them. Here, we choose to use loops to combine all data (so do the distance data). (fig 2)



fig.2

Now all the pre-process work is done, the second step is to use Matplotlib to show the figure. Here, we also meet some problems, when we got the figure, the 31 days of data were all overlapped instead of being connected from left to right. So, we chose to add the date and time together to become the x-coordinate of the figure. (fig.2)

Then, we got the figure of the Earth-Moon distance and height of tides. (fig.3, fig.4)

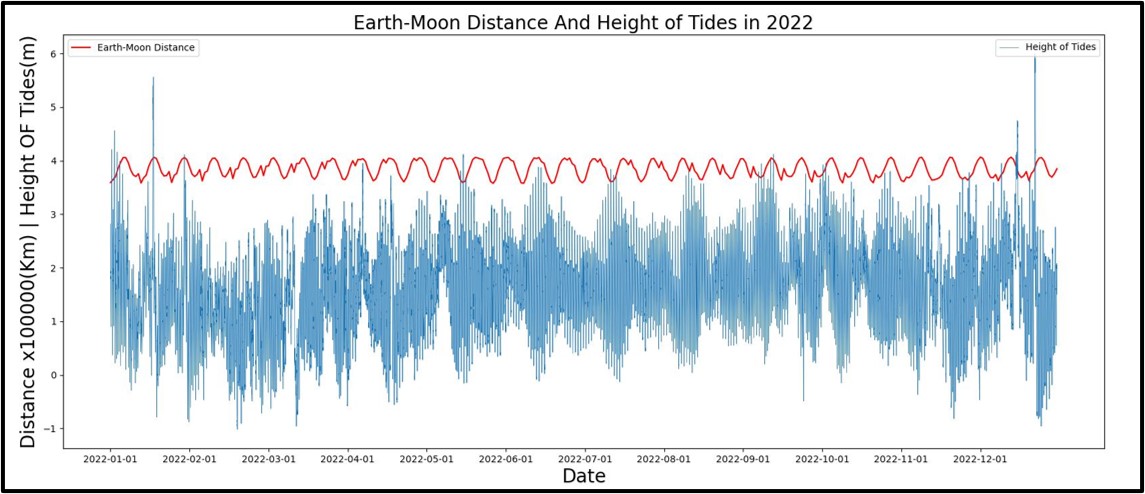


fig.3



fig.4

The datasets of Earth-Moon distance which have been properly cleaned in the prior stages would not be discussed in this section. The original tidal data contains 86700 objects, recording water levels by seconds in the year 2022. The massive amount of data is hard to predict by the emerging models as it includes inappropriate data types, redundancy, and tremendous outliers. The visualization of the whole data (fig.3) does not support accurate observation and analysis. So, the major goal is to clean up the data and increase the scale to an extent for yearly trend analysis.

There are three major steps in data cleaning: adjusting data, monthly average, and sliding average. In the first place, the addition of data points in the same is performed. Daily, mean water level is then obtained by taking the average of summation. Meanwhile, the time data is formatted into date-time datatype in Python corresponding to dates of the daily water level. To further increase the scale of data, the sliding window algorithm is acted on the daily data. To be more specific, the algorithm sums up all water levels between the start and end pointers in the array and takes the mean of the sum. The generated data point is appended into a new array. The distance between the two pointers is called window size which could be adjusted as needed. In this case, the window size is chosen as 10 to avoid overgeneralizing. Consequently, a clean data set is attained for further analysis.

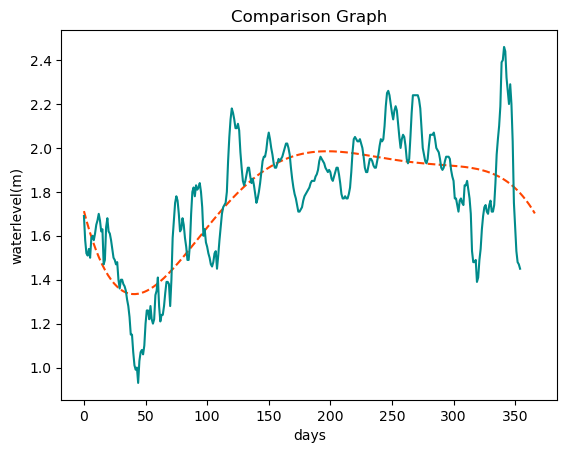
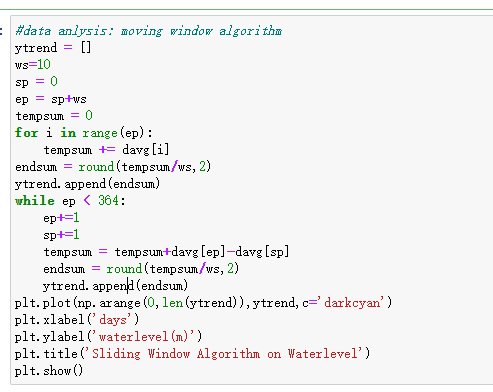


Fig.5 + fig.6

1. Data processing

During data processing, key trends of data are extracted by multiple methods for further analysis.

1. **Polynomial Regression on Tides Data**

Regarding the change in tidal data, polynomial regression is considered to fit in a curve for pattern recognition. A degree of five is adopted for its balance between fitting and over-fitting (Fig.6).

1. **Multiple Seasonal-Trend decomposition using LOESS (MSTL)**

Earth-Moon distance data is processed by MSTL for the repeating trend shown in (fig.7) The trend of data is extracted to a larger scale for better observation (Fig.8).

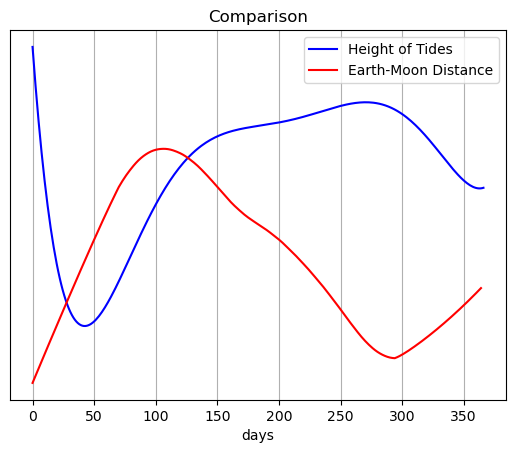
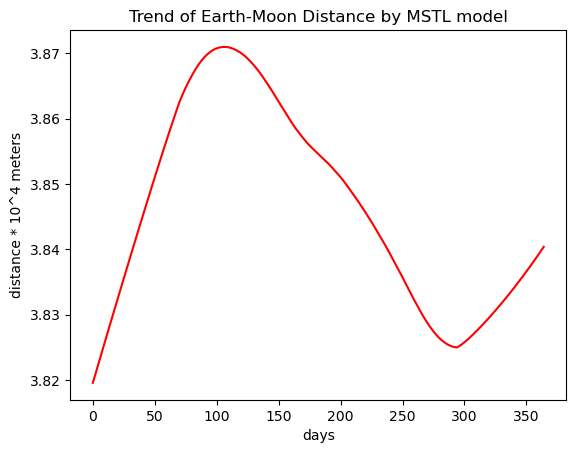


fig.7 + fig.8

1. Data Analysis

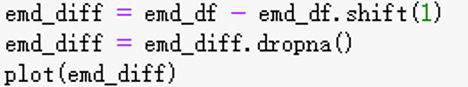
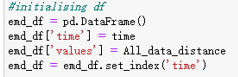
To clearly observe the results of two types of data, Fig.6 and Fig. 7 are separately enlarged and plotted together for corresponding dates on the x-axis (Fig.8). This means that only the trend of the two curves could offer indications whereas the numeric value for Y axis has no specific meaning.

Both curves demonstrate a trend of first climbing and then declining which is similar to half the period of a normal sine-curve. Specifically, the plummeting at the start of the blue line (water level) is interpreted as the end dropping of the last cycle. In addition, a lagging effect could be concluded from Fig.8 since they displayed similar trends. The trend change in the blue line (water level) emerges after a similar change occurred for the red line (Earth-Moon Distance). For instance, fifty days after the red line started increasing, the blue line commenced increasing as well. However, the length of lagging cannot be determined by this single graph. Although both curves illustrated a downward trend at the righter half of the graph, the difference of occurrence in this similarity far outweighs 50 days' length. Further research on this lagging length is expected to be done in the research paper later this summer by our group.

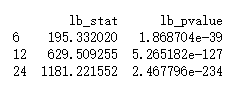
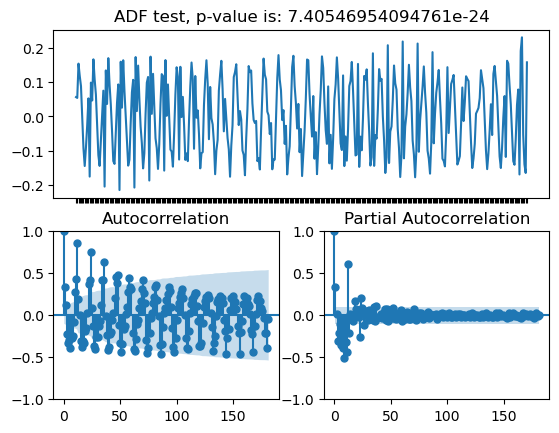
1. Data Prediction
   1. Auto–SARIMA model on Earth-Moon Distance Prediction

The selection of Auto - the SARIMA model is due to the characteristics shown by the Earth-Moon Distance data. Compared to the fluctuating data of tides in Fig. 3, the red line of Earth-Moon Distance shows strong stability and obvious seasonality by months overall, which is in accord with the SARIMA model. Meanwhile, the ‘pmdarima’ package that automatically searched suitable hyper-parameters for the SARIMA model is used here to replace the grid-research method which contains nested loops and consumed a tremendous amount of time. Thus the efficiency of the program is increased significantly.

To successfully conduct Auto - the SARIMA model, the Adfuller test (ADF) is conducted to verify its stability, displaying negative results with p-value = 0.771 greater than standard 0.05. Therefore, the first difference, which is adequate, is applied to the set of data for stabilization. The resulting data set passed the Adfuller tests with a p-value small enough to be ignored. Furthermore, the white noise test (Ljung - Box test) is operated to ensure the data does not exist at random and it passed the test with its p-value at 6 lags already small enough to be ignored. Consequently, we have proved the data to be predicted is stable with some patterns. Important steps are shown in Fig.10.



( Initializing data frame) (First difference step)



(Passed results on ADF) (Passed results on white noise test)

fig.9

After testing the data set, Auto - the SARIMA model is applied with 80% of data for training and 20% for predicting. The RMSE value, which represents the error rate, is calculated as about 0.1. Turning to the details, the fitted line starts with a straight line because the differentiated data has lost the first term, which is in the range of acceptable loss. It can be observed that the fitted line almost overlapped with the trained data. In contrast, the forecast line exhibits some conflicts with the raw data, concentrated at the end of the year time. These conflicts attribute to the flaw of auto-searching hyper-parameters as well as the incompetency of taking means when cleaning data which omits too many data points. Overall, an accurate prediction is made by the model.

* 1. Prophet model on Tides data

Prophet model is established instead of the Auto-SARIMA model which does not sit in the data set. Specifically, the water-level data of tides illustrated both daily and yearly trends while the SARIMA model can easily be disturbed by these characteristics (Fig.10). However, the Prophet model has taken into account the effect of multiple seasonality (Fig.11). The hyper-parameter of Prophet model is set to minimize the influence of yearly seasonality as the annual change is very minimal. As a result, the prediction demonstrates a stable trend with continuous seasonality.

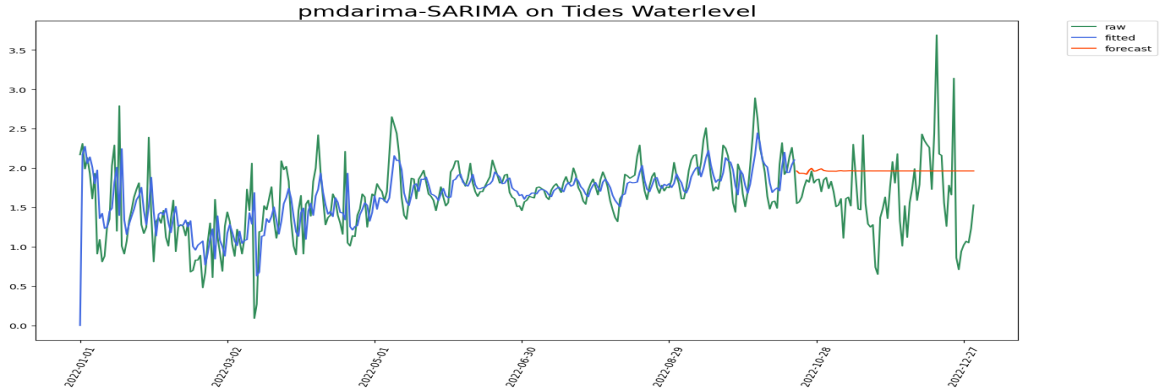


fig 10

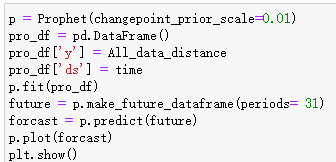
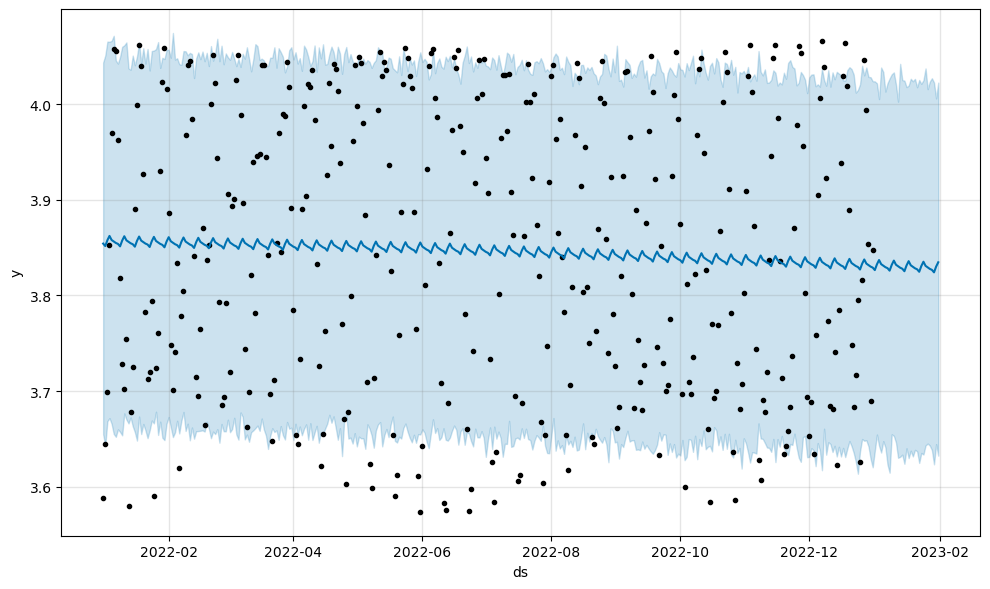


fig.11