Project S (Early): Data Cleaning

Eric Tang
UC Berkeley
Berkeley, CA, 94704
erictang000@berkeley.edu

Calvin Grewal
UC Berkeley
Berkeley, CA, 94704
calvingrewal@berkeley.edu

Abstract

We obtained 20 years of positional data as well as astrological event data (Eclipses, Equinoxes, etc.) for each of the planets in the Solar System, relative to 1000 different locations on Earth for the purpose of being able to train a geocentric model that would be able to predict location of planets relative to time and locations on earth. We then cleaned and organized the positions/time/location data such that each is accessible easily via a pandas dataframe.

1 Introduction

Our dataset consists of 20 years of daily position data and associated astrological event data for each planet in our solar system as observed from 1000 different locations on Earth. We were inspired to make this dataset in order to use Ancient Greek models of astronomy, so we generated our data primarily in a geocentric way that would most reflect models of that time period.

First, since data formatted in an absolute coordinate system such as an Equatorial Coordinate System or a Geocentric XYZ coordinate system implicitly encodes distance data, which was not accessible to ancient observers like Ptolemy or Hipparchus, we decided to also include data set in the Horizontal Coordinate System, which represents planet locations relative to the location of the observer using azimuth and altitude measurements, which are relative to some arbitrary horizon and are more similar to angular measurements that ancient observers may have used in their construction of the mathematical models of the night sky. Since the horizontal coordinate system is not consistent across different locations at any given time, we included the data for 1000 different locations so that users of the dataset would be able to use the whichever location(s) best suit their needs. In addition, we would easily be able to compare the differences between how models generalize when trained across a variety of different locations versus how they would appear when trained just across data from one location, as would have been the case for ancient observers.

The event data consists of various astronomical events including Eclipses, Conjunctions and Oppositions, as well as Solstice and Equinoxes. As opposed to eclipses which in our data refer to only lunar and solar eclipses, conjunctions and oppositions are when planets "line up" with the sun. For example, a Saturn Conjunction is when the Earth is lined up with Saturn such that the Sun is directly in between. We decided add this events data to our position data since Ancient Greeks could see these events and believed them to be prophetic, and may very well have been able to incorporate these events into their astronomical models. One thing to note is that conjunctions for Venus and Mercury are further separated into Superior/Inferior conjunctions, corresponding to when Venus/Mercury is on the opposite or same side of the sun, respectively.

Both the positional data we obtained, as well as the event datasets were combined in order to create our final dataset to be used. Therefore, the final dataset contains position data for each planet every

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

day from 2000-2020, along with any of our selected astrological events happening on any given day. As compared to a dataset which only contains position data, this combined dataset will allow users to both look at how the events are related to position, and even use position data to predict when these events will happen.

2 Methods

For our dataset we primarily used two data sources: JPL's Horizons Systems Ephemerus database and NASA SkyEvents Calendar Database. We also supplemented the JPL Horizons database with additional positional data obtained via the get_body function of Astropy, which given a time, planet, and earth location, is able to return the coordinates of the given planet centered at the given earth location at the given time using a GCRS reference frame.

The first step in our process was data collection. For obtaining data from the Horizons system, we adapted a Perl script for sending HTTP requests to the server that was listed on the website, and called that script from a Python script which we used to change the parameters for the different http requests. Through this process we were able to obtain over 7.5 million lines of positional data corresponding to the 1000 different locations and the 10 different planets over 20 years (starting at Jan 01 2000 and going to Dec 25 2020 with a step size of 1 day). Once we collected the data we then needed to process it and clean it for use in a machine learning context. We decided to place the data into a pandas dataframe for ease of storage and querying for future users. To do this we parsed each of the 10000 text files outputted from the data collection step into strings separated by commas, and then parsed these strings to output the 7 million rows of data into our dataframe.

With the data parsed out into a dataframe, we then proceeded to convert the raw data into more useful values. The Horizons data for each planet came in the Equatorial Coordinate system, but since we wanted data as observed from a location on earth, we used the Astropy library to convert the data into the Horizontal Coordinate system and a 3D vector coordinate system centered at the various Earth locations included in the dataset. This was done using the latitude and longitudinal data obtained for the each of the 1000 locations from the Horizons system. In addition, we added functionality for thresholding visibility of planets based on the value of the apparent magnitude, to allow for users to eliminate observations that may not have been available to ancient observers. This step opens the door to comparison between models trained on a full range of astronomical data vs models trained on more restricted and noisy data. Our final pandas dataframe consisted of the position of the planet of interest in ra/dec, az/alt, and geocentric XYZ coordinates in AU units, as well as the earth location in lat/lon and geocentric XYZ coordinates in geocentric AU units.

For the SkyEvents data, there was no official API or way to download the data, the only way to view data was year by year through the website, so we manually went and copy and pasted 21 years worth of SkyEvents from the website into Excel and then exported it into a CSV from there in order to get it into Pandas. The data was formatted like a calendar with a date associated with an event, e.g. "October 25 2020 Mercury Inferior Conjunction". In order to turn this into usable features, we decided to one-hot encode the data. Given that our position data is daily, meaning we have a row for each day of the 20 year span, it made sense to include each event as a feature, with the corresponding row containing a '1' in the event feature column if the event occurred on that given day.

In order to standardize both datasets and make sure all the dates were consistent, we worked in UTC time and converted everything to a Pandas datetime object, and then merged the two datasets based on the date. This prevented us from running into any time zone issues.

3 Results

We were able to successfully accomplish our goal of creating a combined dataset of positions and events. Our dataset includes three different types of coordinates for each row (ra/dec, alt/az, and geocentric XYZ), along with various visibility metrics (apparent magnitude, surface brightness, etc.), and then the events data.

Rather than just giving the huge dataset by itself, we have included a function which lets users custom query data, with parameters for number of earth locations, number of days, frequency of data, and which planets to return. For example, using this function a user would be able to get 365 days worth of data for Mercury and Venus at a frequency of every 5 days. This is incredibly useful as it makes working with the data much easier.

After generating our data, we did some exploratory analysis to validate our data and check that everything made sense. To do this, we plotted both azimuth and declination vs date for both Mercury and Venus. We can clearly observe a sinusoidal pattern for both plots. In addition, we observe that compared to Venus, Mercury moves at a higher frequency which makes sense as Mercury has a much faster orbit than Venus.

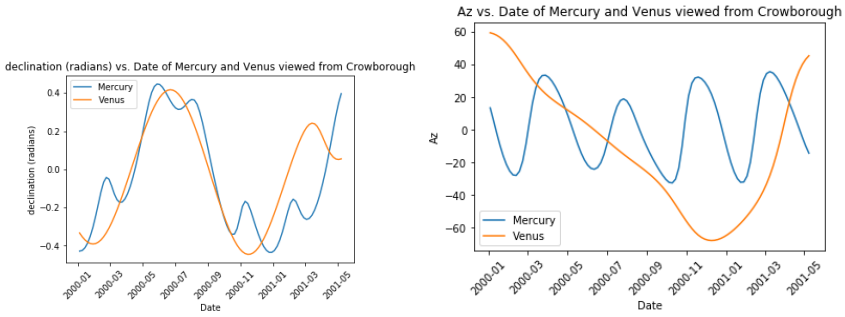

Figure 1: Plots of declination and azimuth vs time on obtained data

	time	ra	dec	Planet	CenterName	CenterLon	CenterLat	az	alt	dist (center to planet AU)	...	Partial Solar Eclipse	Annular Solar Eclipse
0	2000-Jan-01 00:00	18 04 55.50	-24 22 51.5	Mercury	Crowborough	0.15420000	51.0518512	17.74343784755384 deg2	-62.460788027094935 deg2	1.4130914988961565 AU	...	0.0	0.0
1	2000-Jan-02 00:00	18 11 45.95	-24 27 21.3	Mercury	Crowborough	0.15420000	51.0518512	16.374140650186238 deg2	-62.66790959130999 deg2	1.4177089536661651 AU	...	0.0	0.0
2	2000-Jan-03 00:00	18 18 38.19	-24 30 32.3	Mercury	Crowborough	0.15420000	51.0518512	14.965991636726711 deg2	-62.84443967309148 deg2	1.421780515121529 AU	...	0.0	0.0
3	2000-Jan-04 00:00	18 25 32.12	-24 32 23.5	Mercury	Crowborough	0.15420000	51.0518512	13.522132858462863 deg2	-62.98931221027721 deg2	1.4253081010918933 AU	...	0.0	0.0
4	2000-Jan-05 00:00	18 32 27.62	-24 32 53.5	Mercury	Crowborough	0.15420000	51.0518512	12.046232612115194 deg2	-63.101384408765874 deg2	1.4282926192572525 AU	...	0.0	0.0

Figure 2: Some sample rows of our combined dataframe

4 Conclusion

Using our data, teams should be able peform a variety of modeling tasks. We provide the data in a format this is flexible and easy to use for many different types of models. Although the Horizons/SkyEvent data was complete in the sense that there are not any missing data points, we reformatted the data such that it is much more convenient.

With our added features and wrapper function for querying specific data, we hope that many groups will find our data useful. For our final project, we will look to use this dataset to create models based off of Ptolemy's and are confident this data will useful in doing so.

For groups using our data, one thing to note is that some of the events occur infrequently, on the order of once a year, which is something to keep in mind while modeling with the data. This is not something we could get around, since the farther planets take longer to orbit, which leads to events such as conjunction and opposition happening infrequently. Other than that, it should be straightforward to follow the setup and use the dataset.

References

[1] SkyEvent Data: https://eclipse.gsfc.nasa.gov/SKYCAL/SKYCAL.html

[2] HORIZONS Data: https://ssd.jpl.nasa.gov/horizons.cgitop

Link to our project: https://github.com/erictang000/astro-data