Title:

Predictive data modeling of a variable binary star's brightness over a period of time using astrostatistics

Research question:

How can the observable brightness trend of Theta Aurigae, a variable double star, be predicted over a period of time using astrostatistics?

Subject:

Physics

Word count:

3998

Table of contents

Table of contents **2**

Introduction **4**

Research question **6**

How can the observable brightness trend of Theta Aurigae, a variable double star, be predicted over a period of time using astrostatistics? 6

Background Information **7**

Stellar properties 7

Binary systems 9

Stellar variability 9

Regression 10

Methodology **13**

Slooh 13

Star chosen for observation 14

Variables 15

Analysis procedure 16

Data collected **17**

Analysis **19**

Plots 19

Polynomial model 21

Trigonometric model 27

Comparison of results with other pre existing trends 32

Conclusion **34**

Appendix **36**

Appendix i. - The python code for polynomial regression of the app. mag data collected: 36
Appendix ii. - The python code for trig regression of the data: 37

Bibliography **38**

Introduction

When one looks up at the sky, one sees the celestial heavens littered with dots of light. These dots represent many things, to different people. To some, they represent balls of nuclear fusion in hydrostatic equilibrium, eternally fighting the battle between gravity and electromagnetic pressure. To others, they represent the ones that they've lost, comforted by the fact that the ones they loved are gazing down at them from the heavens. But no matter what they mean, stars have always played the part of silent watchers in the evolution of life on earth, from being ever-present beacons guiding the sailors of the old, to being objects of great study to modern day astronomers, providing valuable insight into the workings of the universe.

The brightness of a star is a measure of the power emitted by the star as seen from a distance - the distance between the star and the earth. It is one of the many properties intrinsic to a star, due to the very fact that it radiates electromagnetic radiation. The brightness of a star varies based on the distance and follows the Inverse Square Law[1]. There exist certain star systems which have two stars orbiting around each other called binary star systems. There are also some stars that vary in brightness over time due to certain stellar physical mechanisms. In this Extended Essay, the researcher will be analyzing and modeling the brightness trends of an Alpha2 Canum Venaticorum variable binary star system over a period of 30 days, give or take. Research on Alpha2 Canum Venaticorum variables is quite few and far between, thus this investigation has

---

[1] "Inverse Square Law, General." *Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/Forces/isq.html. Acc. Dec., 2017.

presented the opportunity to carry out research that is relatively rare, since A2 CV variable stars, and specifically the star chosen for this investigation, has not been researched much.

The interest in this topic is due to a personal passion for astrophysics. The researcher always had a desire to conduct an observational experiment in the field of astrophysics in order to observe a mathematical trend in some astrophysical property or the other. Through this investigation, the IB Extended Essay has presented the opportunity to do so.

The application of such kind of modeling of the brightness of stars is quite common, and is the modern way of obtaining the brightness trends of stars. On obtaining the initial observation data for an astronomical object, one can fit a data model that can be used to predict future observations. If the model does not agree with future predictions, it is retrainted/tuned with data until it does so. This cycle repeats and the astrophysical models that we have devised get more and more accurate.

Research question

How can the observable brightness trend of Theta Aurigae, a variable double star, be predicted over a period of time using astrostatistics?

Background Information

Stellar properties

Stars are approximated to be black bodies[2], which are perfect absorbers and emitters of electromagnetic radiation. This approximation means there are a certain set of laws that can be applied to them which can be used to determine their properties and behavior. For instance, Wien's displacement law[3] relates the wavelength of the electromagnetic emission of black bodies (stars in this case) to their temperature. Similarly, the Stefan-Boltzmann law[4] is used to deduce the power emitted by black bodies, and mathematically states that the power emitted by a black body is related to its surface area (A) and temperature (T) as follows:

$$P = \sigma A T^4$$

Where σ is the Stefan-Boltzmann constant having a value of

$$5.6703 \times 10^{-8} \frac{W}{m^2 K^4}$$

For stars, which are assumed to be spherical in shape (so $A = 4\pi R^2$), the power emitted is called Luminosity and is given by the same law as:

$$L = \sigma 4\pi R^2 T^4$$

Where R is the radius of the star.

---

[2] "Blackbody Radiation | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/b/blackbody radiation. Acc. 19th December., 2017.
[3] "Star Temperatures." *Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/wien.html. Acc. 18th December., 2017
[4] "Stefan-Boltzmann law."*Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/thermo/stefan.html. Acc. 18th December., 2017

From luminosity, we get the definition of the brightness[5] of a star, which is the luminosity received per unit area at a specific distance from the star. Unlike luminosity, brightness is a subjective quantity and varies depending on the distance, changing in accordance with the inverse square law[6]. The brightness is given in relation to luminosity as:

$$b = \frac{L}{4\pi d^2}$$

Where d is the distance to the star in meters. The brightness of a star can also be expressed in terms of a logarithmic magnitude scale. This scale is of two types - absolute magnitude and apparent magnitude.

The apparent magnitude of a star is the logarithmic measure of its brightness as seen from Earth whereas absolute magnitude is the logarithmic measure of the star's brightness as seen from a distance of 10 parsecs. Being logarithmic, a more negative value on the scale means higher brightness. The relation between the distance (d, in parsecs), absolute magnitude (M), and apparent magnitude (m) is given as:[7]

$$m - M = 5\log(\frac{d}{10})$$

The brightness of a star can be obtained through flux photometry, which uses telescopes to collect light from the star and determine brightness from the amount of light incident.

[5] "Luminosity and How Far Away Things Are." *European Imperialism*, pages.uoregon.edu/soper/Light/luminosity.html. Acc. 3rd February., 2018
[6] "Inverse Square Law, General." *Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/Forces/isq.html. Acc. 19th December., 2018
[7] "Absolute Magnitude | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/A/Absolute Magnitude. Acc. 12th January., 2018

Binary systems

Binary star systems are systems in which two gravitationally-bound stars orbit each other over time. More specifically, the stars orbit their common center of gravity called the barycenter.[8] Binary orbits decay over time because the orbit is not spherically asymmetric so it radiates energy in the form of gravitational waves. However, for systems where the objects in question are not massive, exotic bodies, the decay rate is extremely miniscule and any significant change in the orbital radius and other properties can only be noticed after millions of years.

Binaries are classified into five different types[6]: Optical double, Visual binary, Spectroscopic binary, Eclipsing binary, and Astrometric binary. Binary systems are useful for astronomers because they are one of the easiest observable cases of how gravity interacts with the other three forces of nature.

Stellar variability

There exist certain stars that vary in brightness over time. This variation is caused by some physical mechanism that affects one or more of the star's properties that are linked to its luminosity. Variable stars are unique in that all variables have similar properties and there's very little diversity in a group of variables. They have also been important throughout the history of

---

[8] "Lecture 13: Binary Star Systems, Masses of Stars." *Culture, Creating Change, Physical Activity for All | Alberta Centre for Active Living*, sites.ualberta.ca/~pogosyan/teaching/ASTRO_122/lect13/lecture13.html. Acc. 4th March., 2018

space sciences, and are a core component in a large number of astronomical methods that look to deduce properties about the universe. An example of a popular variable star group is the group of Cepheid variables[9].

As far as this investigation is concerned, there's a certain variable star group called Alpha2 Canum Venaticorum variables[10]. A2 CV variables are chemically peculiar main sequence stars that possess unusually powerful magnetic fields and their light spectrum contains silicon, strontium, and chromium spectral lines. These spectral lines fluctuate along with the magnetic fields. The periods of these variations are the same and equal the star's rotation period. Furthermore, the magnetic fields cause a varied distribution of metals in the atmosphere of these stars so their surfaces vary in brightness from point to point, giving rise to their variable nature. The class name "Alpha2 Canum Venaticorum" comes from the star α^2 Canum Venaticorum, which is a star in the binary system of Cor Caroli located in Canes Venatici and the first variable of this type to be identified.

Regression

Regression analysis is a field of statistics that is used to predict the trend of future data based on the analysis of existing data by estimating the relationship observed between two or more variables. Regression analysis talks about modeling the relationship between how the value of one variable changes with respect to another. The most introductory form of regression is called

[9] "Cepheid Variable Stars." *Density Definition*, astro.unl.edu/naap/distance/cepheids.html. Acc. 29th September., 2018
[10] Revolvy, LLC. "'Alpha2 Canum Venaticorum Variable' on Revolvy.com." *Trivia Quizzes*, www.revolvy.com/page/Alpha2-Canum-Venaticorum-variable. Acc. 8th October., 2018

linear regression, where there is a linear relationship between the variables being modeled, and the process involves finding the 'line of best fit' by minimizing the deviation of the line of best fit from the actual data points through trial and error.

Apart from linear regression, there are many other types of regression, mainly:

- **Polynomial regression**: Polynomial regression[11], is a special case of linear regression where the variables in question do not follow a linear relationship but instead an nth-degree polynomial curve must be modeled to fit their trend. Polynomial regression mainly involves turning the coefficients of this nth-degree polynomial to fit the data with the maximum accuracy possible.
- **Trigonometric regression**: Trigonometric regression[12], is regression in which the variables follow a trend that obeys some trigonometric function at its heart. This trigonometric function is accompanied by some constant coefficients that, when tuned, best fit the data points in question. When that function is the trigonometric sin, it is called sinusoidal regression.

All kinds of regression involve calculating a loss function for the model being used, and reducing this loss. The loss function determines how much the data points deviate from the curve being modeled for them, and the goal of regression is to then minimize function to make the model a

[11] "What Is Linear Regression?" *Statistics Solutions*, www.statisticssolutions.com/what-is-linear-regression/. Acc. 10th August., 2018
[12] Roberts, Donna, and Frederick Roberts. *Rotation - MathBitsNotebook(Geo - CCSS Math)*, mathbitsnotebook.com/Algebra2/Statistics/STregression2.html. Acc. 10th August., 2018

more accurate fit. Such minimization is done through another process called gradient descent - which is finding the minima of the loss function.

In data science, there are two certain phenomena that are prevalent in stochastic modeling that cause problems in accuracy, namely overfitting and underfitting.

Overfitting[13] is the process of tuning a model to the extent that it fits the data too well - which, although might seem desirable, isn't beneficial because overfitted models will predict a trend that have a high probability of significantly deviating from any possible observable data obtained in the future. This phenomenon is prevalent because almost all data trends are stochastic to some extent, which sets a limit on the maximum accuracy a model can take. Overfitting can be minimized by making sure that one does not train the model too much/multiple times with the same data due to which the model "memorizes" the data instead of learning from it.

Underfitting[14] is the process of tuning a model too poorly or not tuning a model enough, making the model a bad fit for the data. Underfitting is obviously not beneficial to any predictive data model. In the case of polynomial regression, underfitting is when the model curve is of a lesser degree of power than the one that is required to accurately fit the model. Underfitting can be minimized by training the model just enough (too much would lead to overfitting) with a variegated dataset.

---

[13] Kenton, Will. "Overfitting." *Investopedia*, Investopedia, 13 Dec. 2018, www.investopedia.com/terms/o/overfitting.asp. Acc. 5th September., 2018
[14] "What Is Underfitting | DataRobot Artificial Intelligence Wiki." *DataRobot*, 13 Dec. 2018, www.datarobot.com/wiki/underfitting. Acc. 6th September., 2018

Underfitting is relatively more common than overfitting, mainly because all models exhibit underfitting initially before being trained with data over multiple epochs to develop into good predictors.. A good training procedure is one that finds a balance between underfitting and overfitting, so as to not allow either to happen.

Methodology

Slooh

The methodology involves collecting data on the star's brightness/apparent magnitude values over a period of 30 days or so. The non possession of a telescope of any kind and the location being a city where light pollution is significant enough to hinder traditional astrometry means that the data collection will be done using an online telescope service called Slooh[15], which allows astronomy enthusiasts from around the world to access professional observatories located in remote locations and schedule live observations of stellar objects

Slooh offers access to an interface where one can observe a stellar object live. The Astronomer membership of Slooh provides FITS data and observation logs, which contain parameters such as apparent magnitude and luminosity. An observation will be scheduled every single day at about the same time for a period of 30 days, and the data will be extracted from the observation logs and compiled to be analyzed later on.

[15] "Slooh." *Slooh*, www.slooh.com/. Acc. 17th February., 2018

Star chosen for observation

The chosen star is designated as Mahasim or Theta Aurigae[16] - a spectroscopic binary star system in the constellation of Aurigae. The star is located 166 light years away in the Auriga constellation. Under the Harvard spectral classification[17], the primary companion (Theta Aurigae A) is an A-type star, about three times more massive than the Sun. The secondary companion (Theta Aurigae B) is an F-type main sequence star that is much fainter than the first.

Theta Aurigae A is an Alpha2 Canum Venaticorum variable[18], with a brightness period of 1.37 days. This is relatively short, considering there are variable stars whose pulsation periods are in the order of years. This short pulsation period is what made the Theta Aurigae system ideal for this investigation. The periodic brightness variations is because of the strong magnetic field of the star that is characteristic of an A2 CV variable. With regards to the binary system, Theta Aurigae A and B orbit each other every 200 years or so.

---

[16] "Theta Aurigae." *SIMBAD Astronomical Database - CDS (Strasbourg)*, simbad.u-strasbg.fr/simbad/sim-id?Ident=theta aurigae&NbIdent=1&Radius=2&Radius.unit=arcmin&submit=submit id. Acc. 21st February., 2018

[17] "Harvard Spectral Classification | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/H/Harvard Spectral Classification. Acc. 12th January., 2018

[18] "Mahasim (Theta Aurigae, 37 Aurigae)." *Universe Guide*, www.universeguide.com/star/thetaaurigae. Acc. 24th February., 2018

Variables

The independent variable in this investigation is the date (point in time) at any instant. The dependent variable is the brightness of the star system.

The constant variables are:

- Distance to the star from the Earth.
- The time of observation for each day.
- The magnification of the Slooh telescope.

Investigation procedure

Once the data collection and compilation procedure has been completed, the data points will be plotted on an apparent magnitude vs time period plot and the trend in magnitude values will be observed

Then, a curve of best fit will be modeled for the data points in order to predict the further trend in brightness for the following months. Regression will be used for the modeling procedure, and the regression technique will be implemented using the Python programming language and Desmos - an online service for plotting and analyzing data.

Hypothesis: The initial hypothesis for the data trend is a sinusoidal curve, which will be challenged in this exploration.

To implement regression in python, the following packages will be used:

- Numpy[19] - scientific computation and linear algebra operations package.
- SciPy[20] - scientific computation package for operations such as regression, optimization, etc.
- Matplotlib[21] - plotting and graphing library.

---

19 "NumPy¶." *NumPy - NumPy*, www.numpy.org/. Acc. 21st January., 2019
20 "SciPy.org¶." *SciPy.org - SciPy.org*, www.scipy.org. Acc. 21st January., 2019
21 "Installation¶." *Matplotlib: Python Plotting - Matplotlib 2.2.3 Documentation*, matplotlib.org/. Acc. 21st January., 2019

- Scikit-learn[22] - machine learning and numeric analysis in python.

Data collected

(The data was collected over a period of 21 days)

Date	App. mag (logarithmic)	Luminosity (Watts)
21-10-2018	2.632214144151	1.006763E+29
22-10-2018	3.272355	1.006940E+29
23-10-2018	2.7607268	1.006540E+29
24-10-2018	3.4447693	1.006550E+29
25-10-2018	3.6799583	1.00638E+29
26-10-2018	3.5055826	1.006689E+29
27-10-2018	2.9330038	1.006793E+29
28-10-2018	2.8568107	1.006227E+29
29-10-2018	3.3276575	1.006396E+29

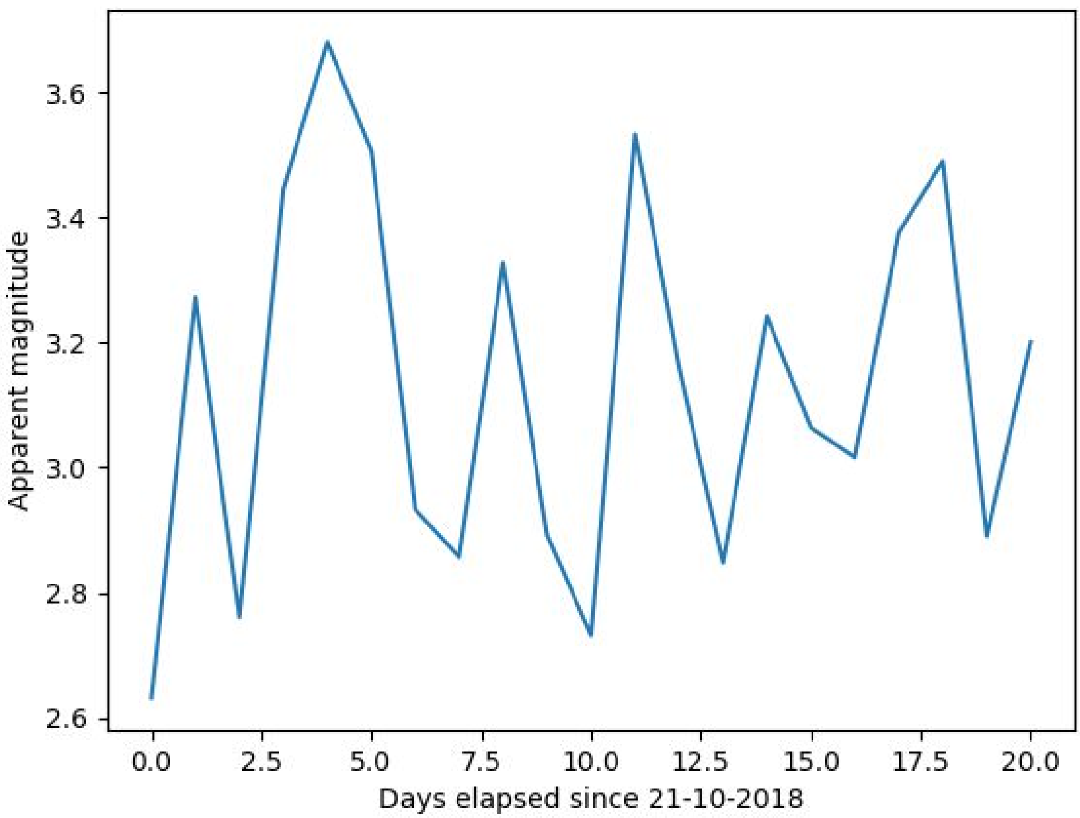
[22] "Scikit-Learn." *1.4. Support Vector Machines - Scikit-Learn 0.19.2 Documentation*, scikit-learn.org/stable/index.html. Acc. 21st January., 2019

30-10-2018	2.8926306	1.00675E+29
31-10-2018	2.7319663	1.006195E+29
1-11-2018	3.5326295	1.006421E+29
2-11-2018	3.160792	1.006119E+29
3-11-2018	2.8481617	1.006925E+29
4-11-2018	3.2420505	1.006898E+29
5-11-2018	3.0637345	1.006798E+29
6-11-2018	3.0162536	1.006439E+29
7-11-2018	3.3754508	1.006308E+29
8-11-2018	3.489457	1.006239E+29
9-11-2018	2.8902778	1.006723E+29
10-11-2018	3.2010884	1.006812E+29

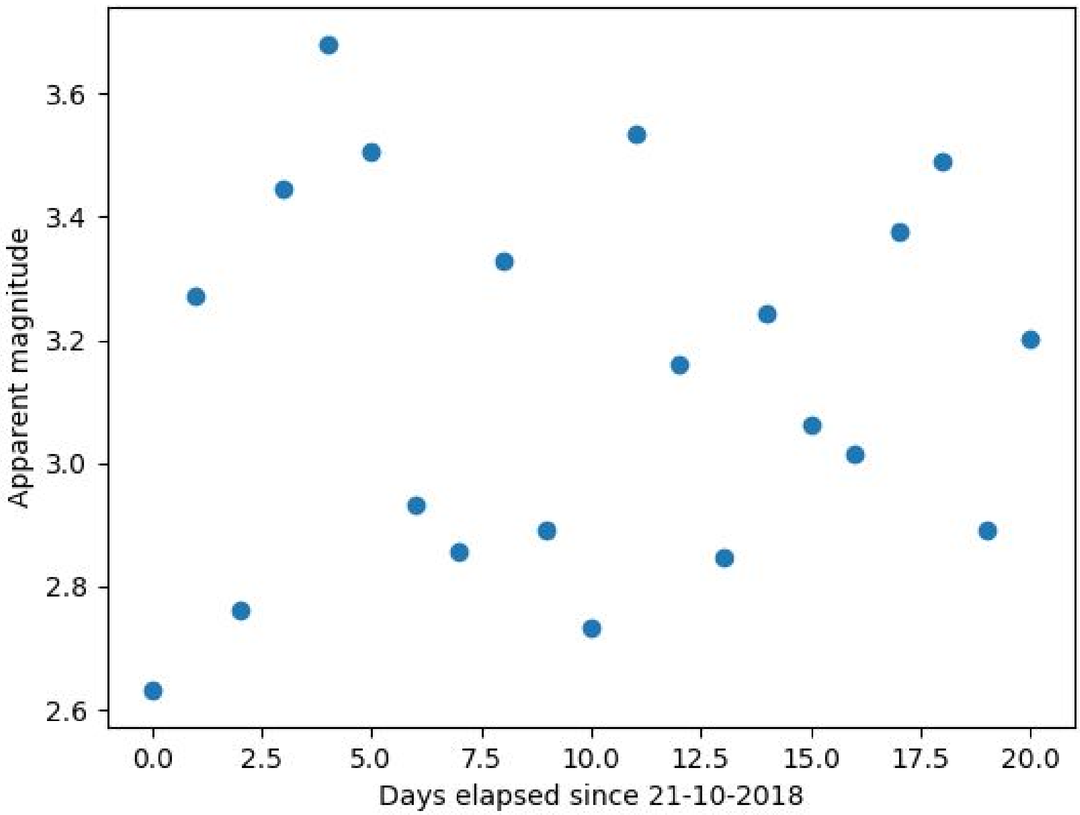
Analysis

Plots

The line plot for the above data is:



And the scatter plot for the data is (x-axis: days elapsed since 21-10-2018, y-axis: app. mag):



Initial thoughts are that the plots appear to show some kind of rough trend in the data. This rough trend seems to be periodic, with the periodicity possibly arising due to the variable nature of Theta Aurigae A. However, the data is also quite irregular in form. which can be explained by considering the fact that no two bright regions on the surface of the star achieve the same peak brightness. Some bright regions may be be brighter than other bright regions which gives a wide range of observed brightness values, or in other words, there is no fixed peak brightness for every bright period, making the collect apparent magnitude values data appear so seemingly random.

However, further analysis can be made by employing regression. Based on the trend, the various models that possibly fit the data points are:

- **An nth degree polynomial model**: The data can be modeled using an nth degree polynomial. The degree of the polynomial to be used can be decided based on the coefficient of determination R^2, which indicates how well the model fits the data. R^2 values range between 0 and 1. A higher value of R^2 means a better fit, but an unusually high R^2 value might imply overfitting rather than a suitable model. The standard form of such a model is:

  $$ax^n + bx^{n-1} + cx^{n-2} + dx^{n-3} + \ldots \epsilon$$

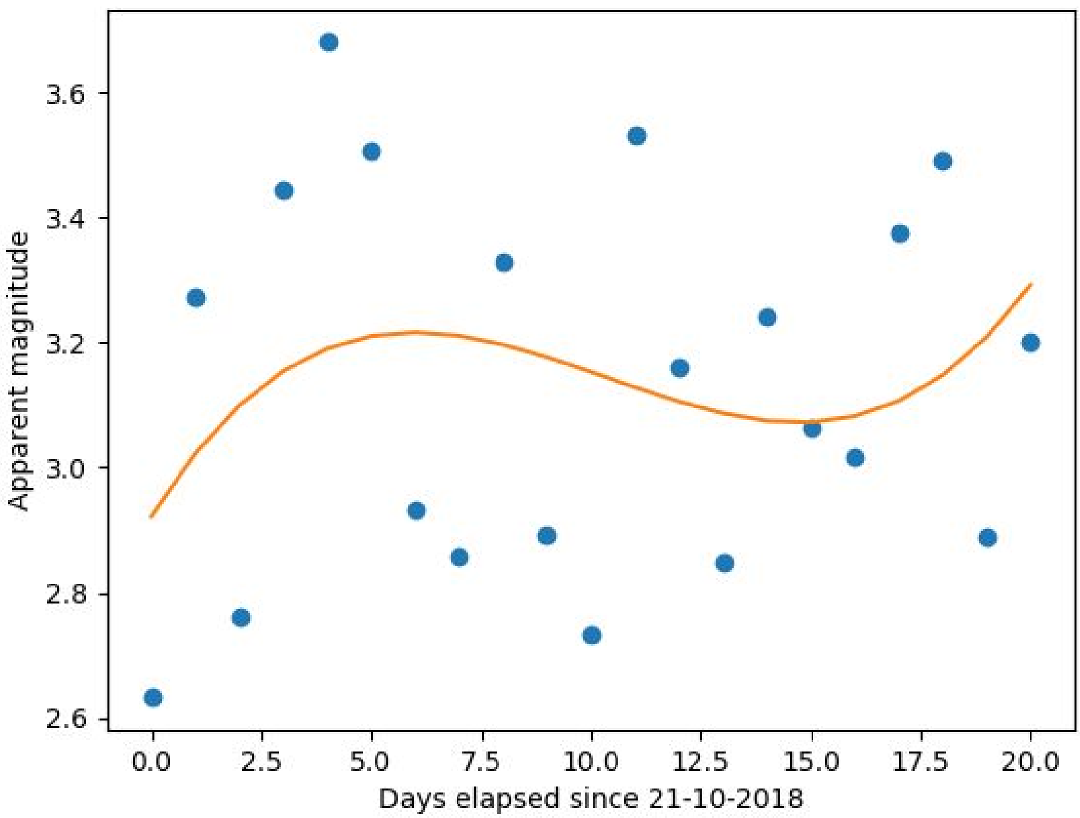
- **A trigonometric model**: The data can be modeled using a trigonometric model. more specifically a sinusoidal regression function. Again, there's an R^2 value even for this model which depends on the tuning constants and must be realistically high for ideal accuracy. The standard form of a sinusoidal model is:

  $$a \sin\left(b(x+c)\right) + d$$

Polynomial model

The polynomial model can be constructed through polynomial regression using an nth degree polynomial. The degree of the polynomial depends on the distribution/apparent trend in the data points. The python code used for polynomial regression of the app. mag data is located in Appendix i.

Since there are only 21 data points, a polynomial of a degree less than 5 won't be very accurate in modeling the data points - there is a rough correlation between the no. of data points and the degree of polynomial needed for modeling. The more no. of data points, the lesser the degree of the modeling polynomial. For instance, the plot below has been generated by using a 3rd-degree polynomial as the model:



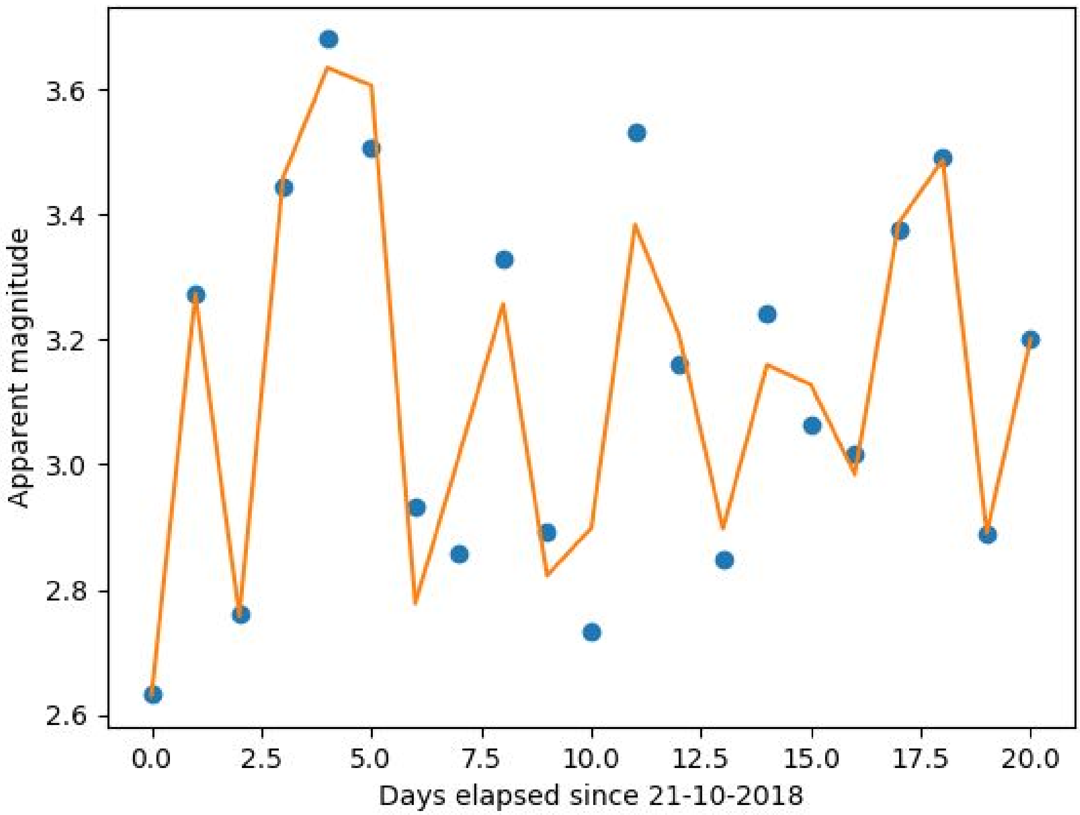
It is clearly evident that the polynomial does not fit the erratic trend of the data points. Therefore, we must go higher in degree, possibly into double digits. After some trial and error with various degrees, the conclusion is that the data can only be accurately represented by polynomials of degree 10 and higher.

Furthermore, on increasing the degree of the polynomial, R^2 increases as well, meaning the accuracy tends to get better. However, there is a limit to this, as beyond a certain degree, overfitting tends to become evident. For this data, the numpy module throws an overfitting warning when the degree exceeds 17:

```
RankWarning: Polyfit may be poorly conditioned
```

This hunch of overfitting is backed by the fact that there is also an anomalous spike in the R^2 value from 17 to degrees higher than 17, with the value going from 0.9248 for 17 to 0.9383 for 18 and then remaining in the 0.97-0.98 range for 19 and higher. Such an anomalous spike is usually caused by overfitting.

Thus, the conclusion is that most accurate model possible is produced by a polynomial degree of 17, below which the data isn't represented well enough by the models (i.e, underfitting), and above which the data is being memorized by the model (i.e overfitting). A 17th degree polynomial model generates the following plot:

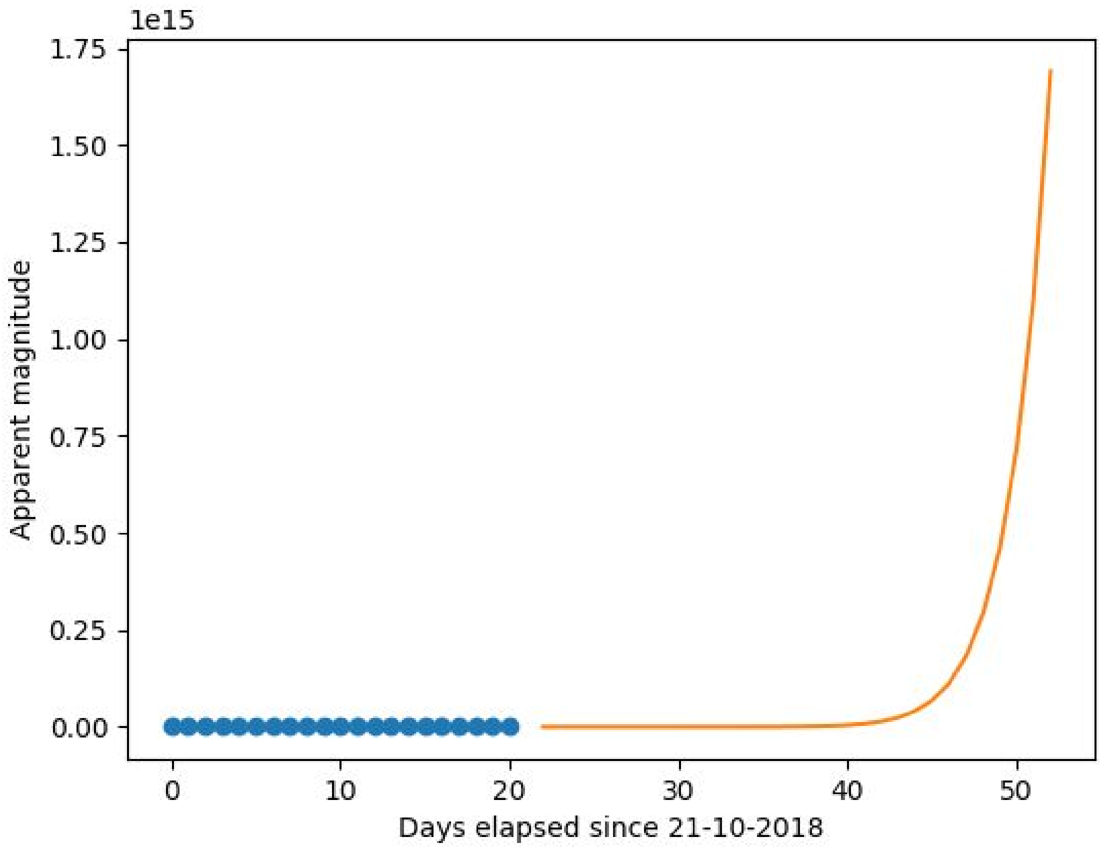


The table below contains the 17 coefficients + 1 constant of the terms of this polynomial model:

Degree/power of the term	Coefficient value
17	$5.76445160 \times 10^{-13}$
16	$-9.98682229 \times 10^{-11}$
15	$7.92645966 \times 10^{-9}$
14	$-3.81870683 \times 10^{-7}$

13	$1.24729328 \times 10^{-5}$
12	$-2.92194125 \times 10^{-4}$
11	$5.06498487 \times 10^{-3}$
10	$-6.60715792 \times 10^{-2}$
9	$6.53259771 \times 10^{-1}$
8	-4.89259171
7	2.75411359×10
6	$-1.14703399 \times 10^{2}$
5	3.44404939×10^{2}
4	$-7.15539607 \times 10^{2}$
3	9.61459471×10^{2}
2	$-7.37630329 \times 10^{2}$
1	2.39409033×10^{2}
0 (constant)	2.63213905

Now, on using this polynomial to predict the brightness trend for the next 30 days (the prediction section in the code in Appendix i), the below plot is obtained:



This trend is clearly not logical, as it shows that over the next month, the apparent magnitude values reach orders of magnitude of 15 and so on, which is physically impossible. Therefore, even though the polynomial model does a good job of fitting the data points, it fails to produce a logical trend. The reason is that the brightness of Theta Aurigae follows a periodic trend, and polynomial models seldom predict a periodic trend (they are actually used to predict exponential trends). Thus, we find ourselves arriving at the second model.

Trigonometric model

The trigonometric model can be constructed using trig regression by using a trigonometric function as the basis. As stated before, the standard trig function that will be used is:

$$y = a\sin\left(b(x+c)\right) + d$$

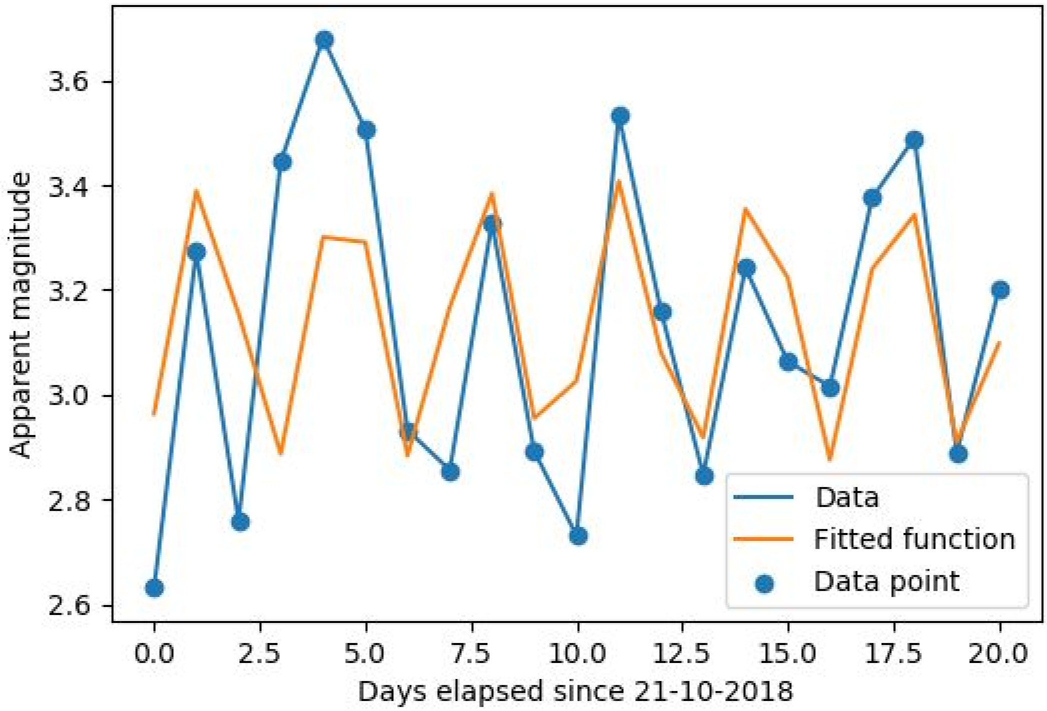
Where the coefficients a, b, c, d will be tuned and assigned appropriate values based on regression. The python code for trig regression of the data is located in Appendix ii.

Trigonometric functions are periodic functions, and do a better job at depicting periodic data such as the apparent magnitude curve over a period of time.

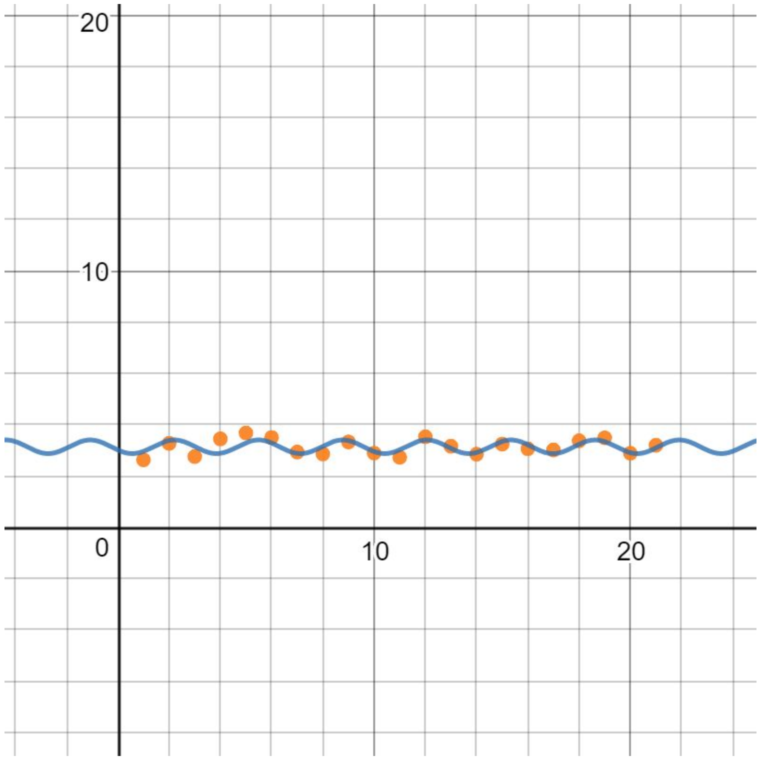
In the code, the sciPy module's optimize function is used to minimize the mean square loss of the curve of best fit to obtain an accurate model of the data. Upon optimization, the R^2 value for this model is 0.397123730200997, which is by no means a good value (a good R^2 value is one that is greater than 0.9). To improve the R^2 value, this investigation needs either:

a. More data to tune the parameters better and obtain a more accurate model, which has a higher R^2
b. A better trigonometric function that is more accurate in depicting the data.

The trigonometric regression model plotted along with the data produces the following plot in Python:



And the below plot in Desmos[23] (x-axis: days elapsed since 21-08-2018, y-axis: app. mag) :



Evidently, this curve of best fit produced from trig regression is significantly less erratic than the curve produced by polynomial regression. Furthermore, this sinusoidal regression involves far less parameters (4) than a 17th degree polynomial (18), and there is less overfitting. The curve also fits the data trend quite well, and has the capability in depicting the periodic nature of the data.

The value of the coefficients a, b, and d produced for this model from Python and Desmos are:

$a = 0.266392353636254$

[23] "Explore Math with Desmos." *Desmos Graphing Calculator*, www.desmos.com/. Acc. 3rd October., 2018

$b = 1.9131934962723058$

$d = 3.1424629041267904$

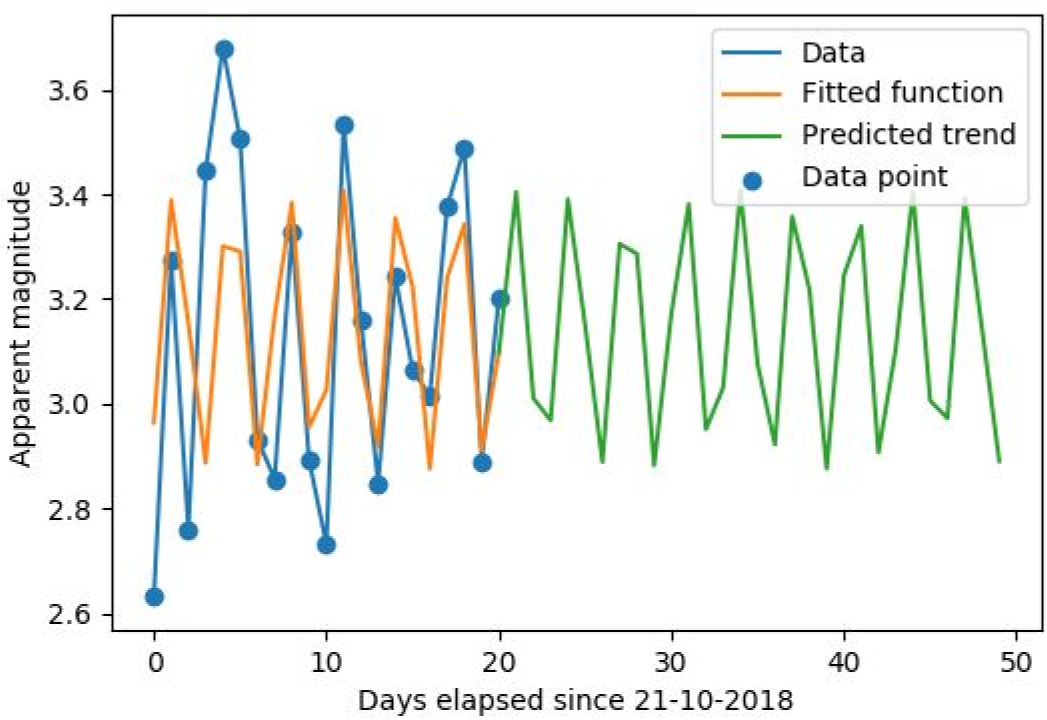
Desmos and Python, however, give different values for parameter c. Desmos gives a value of:

$$c = 11.7535$$

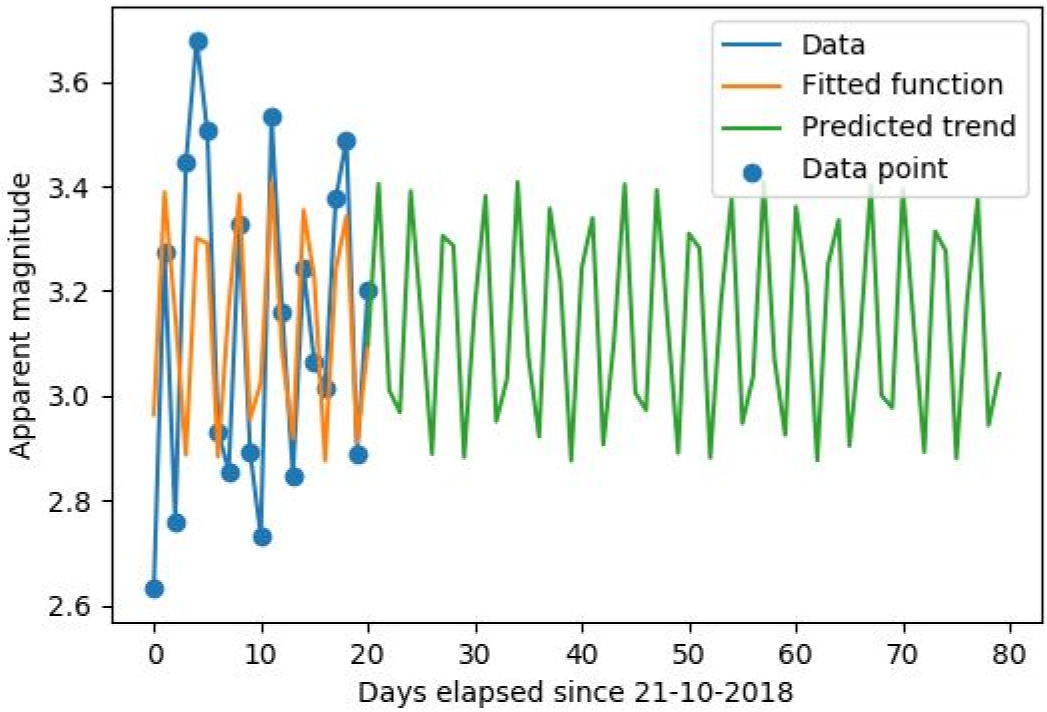
Whereas Python predicts a value of:

$$c = 2.9011427055850487$$

Using this model and Python's c value, the trend for the apparent magnitude of Theta Aurigae over the next 30 days starting from the 11th of November 2018, is predicted to be:



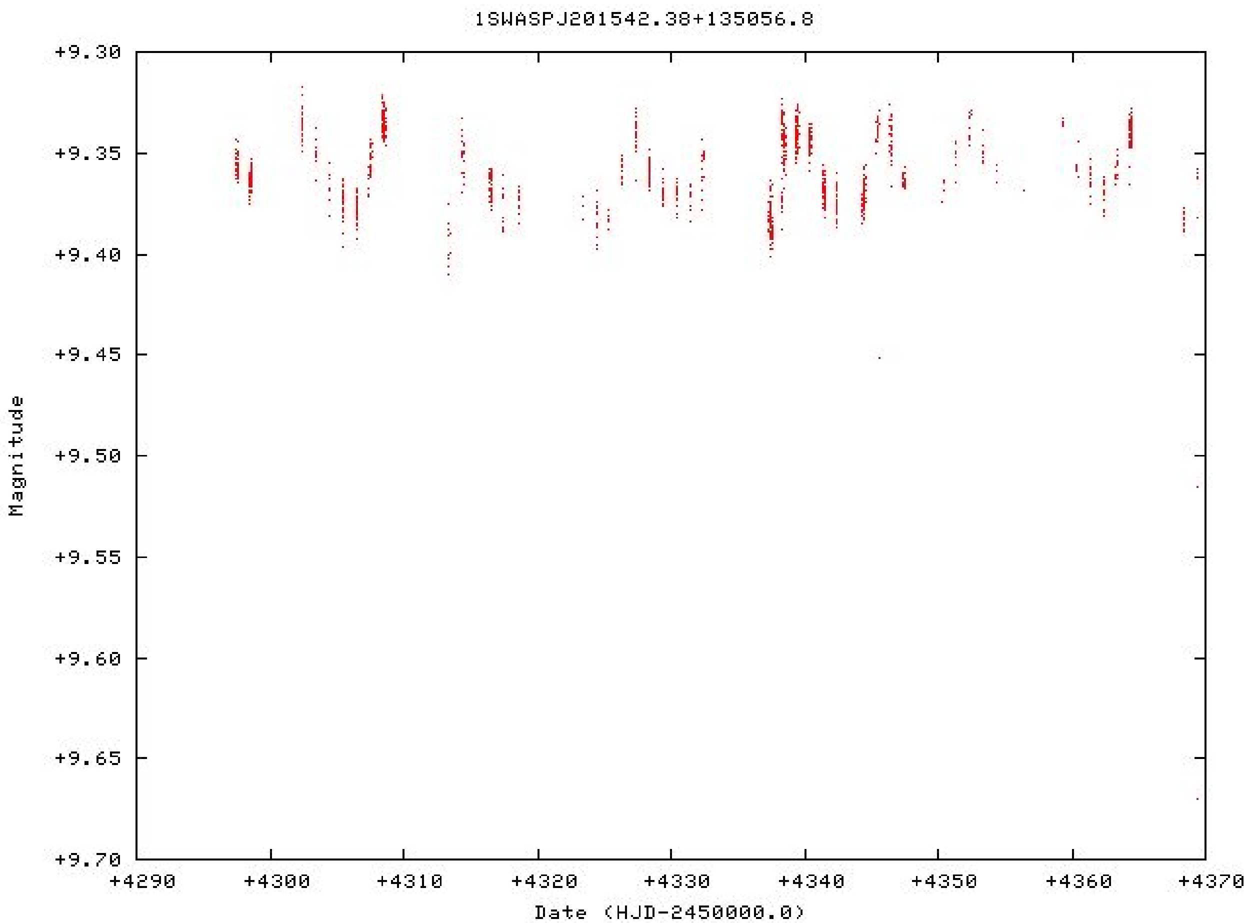
And the brightness trend over the next 60 days starting from 11-18-2018, is predicted to be:



Although the trend wouldn't entirely be consistent with future data, it does show the general direction that the data could take. Furthermore, if more data can be collected, the model can be refined and the parameters can be tuned better to obtain a more accurate prediction, and so on. Therefore, a trigonometric model, apart from fitting the data points, also predicted the trends in apparent brightness for the future.

Comparison of results with other pre existing trends

Given below is the apparent magnitude trend of another Alpha2 Canum Venaticorum variable star[24] called NS Del or 1SWASPJ201542.38+135056.8[25]:



This trend is quite similar to the brightness trend obtained in this investigation for Theta Aurigae. The similarity is that NS Del's trend is periodic in nature and there is no fixed peak brightness

[24] "SuperWASP Observations of Variable Stars." *The Life of a Star*, www.astro.keele.ac.uk/workx/superwasp-variable-stars/Rotating.html. Acc. 27th June., 2018

[25] "NS Del." *SIMBAD Astronomical Database - CDS (Strasbourg)*, simbad.u-strasbg.fr/simbad/sim-basic?Ident=NS Del. Acc. 28th December., 2018

for every bright period/region in the graph. From all this, it can be concluded that all Alpha2 Canum Venaticorum variable stars essentially follow the same kind of trend - one that is periodic in nature with the time period being in the order of days and one that does not have a fixed peak brightness for the bright regions.

Conclusion

Using statistical techniques, even though a model that fits the data exactly could not be obtained (due to the seemingly varied/random nature of the collected data points), it was possible to model and obtain a rough idea about the direction that the brightness of the star will follow over the course of time. Such kind of predictive modeling lies at the heart of astrostatistics and regression-based statistics in general.

It was seen that although a polynomial model was a fit for the data, it failed in making a logical prediction of future trends of brightness. However, the trigonometric model, apart from being a fairly good fit for the data, was competent enough to predict the periodic future trends of brightness. The failure of the polynomial model in the predictive regime was an unexpected and surprising outcome, because it did fit the data quite well. However, from a mathematical perspective, it was expected because polynomial models are characteristic of increasing exponentially, which does not correspond to the periodic nature of binary brightness variations. The success of the trigonometric model meant that the initial hypothesis that the data will follow a sinusoidal trend has turned out to be true.

The main challenge that was faced with this investigation was with regards to the data collection. As mentioned in the Slooh section, not possessing a telescope of any kind and living in a city with high amounts of light pollution made data collection by traditional methods impossible. After much research and searching only was Slooh uncovered, which turned out to be a blessing in disguise. Other challenges include finding the right mathematical model for trigonometric regression and initial difficulty in finding the right degree in polynomial regression.

Regarding uncertainties, Slooh does not publish the uncertainty values associated with their telescopes therefore the systematic uncertainty values are unknown to calculate any possible uncertainties. However, taking an ansatz, even the highest possible uncertainty values do not cause major changes in the brightness data, and the general trend observed will be still be sinusoidal.

A limitation of this investigation is that there wasn't enough data to obtain a more accurate model. Therefore, the model can be refined by collecting data over a longer period - a couple of months or half a year, which would give a more accurate prediction of the brightness trend. Furthermore, on collecting more data, the seemingly random trend will turn out to not actually be random after all, because after one full rotation of the star, apparent magnitude values tend to repeat.

Appendix

Appendix i. - The python code for polynomial regression of the app. mag data collected:

```
# import statements below
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import r2_score

y = np.array([2.632214144151, 3.272355, 2.7607268, 3.4447693, 3.6799583, 3.5055826,
2.9330038, 2.8568107, 3.3276575, 2.8926306, 2.7319663, 3.5326295, 3.160792, 2.8481617,
3.2420505, 3.0637345, 3.0162536, 3.3754508, 3.489457, 2.8902778, 3.2010884]) # y values
x = np.array(list(range(len(y)))) # x values

n = 3  # the degree of the polynomial (change to 17 for 17th degree)

p = np.poly1d(np.polyfit(x, y, n)) # fits the polynomial curve to the data

t = np.array(list(range(len(y)))) # testing data
print("R^2: {}".format(r2_score(y, p(t)))) # displays the R^2 value

plt.plot(x, y, 'o', t, p(t), '-') # plots the data
plt.xlabel('Days elapsed since 21-10-2018') # labels the x-axis
plt.ylabel('Apparent magnitude') # labels the y-axis
plt.show() # generates the plot

'''
# prediction section below (uncomment this section to run the prediction for the next 30
days):

k = np.array(list(range(22, 53)))
plt.plot(x, y, 'o', k, p(k), '-')
plt.xlabel('Days elapsed since 21-10-2018') # labels the x-axis
plt.ylabel('Apparent magnitude') # labels the y-axis
plt.show()
'''
```

Appendix ii. - The python code for trig regression of the data:

```
# import section below
import numpy as np
import matplotlib.pyplot as plt
from scipy import optimize
from sklearn.metrics import r2_score

y = np.array([2.632214144151, 3.272355, 2.7607268, 3.4447693, 3.6799583, 3.5055826,
2.9330038, 2.8568107, 3.3276575, 2.8926306, 2.7319663, 3.5326295, 3.160792, 2.8481617,
3.2420505, 3.0637345, 3.0162536, 3.3754508, 3.489457, 2.8902778, 3.2010884]) # y-values
x = np.array(list(range(len(y)))) # x-values

def trig_func(x, a, b, c, d): # trig modelling function
  return (a * np.sin(b * (x + c))) + d # a, b, c, d are coefficients that must be tuned

params, params_covariance = optimize.curve_fit(trig_func, x, y,
                                p0=[2, 2, 2, 2]) # optimize the curve of best fit

print("R^2: {}".format(r2_score(y, trig_func(x, params[0], params[1], params[2],
params[3])))) # displays the R^2 value


plt.figure(figsize=(6, 4))

plt.scatter(x, y, label='Data point') # scatter plot
plt.plot(x, y, label='Data') # line plot
plt.plot(x, trig_func(x, params[0], params[1], params[2], params[3]),
       label='Fitted function') # figure of best fit
plt.xlabel('Days elapsed since 21-10-2018') # labels x-axis
plt.ylabel('Apparent magnitude') # labels y-axis

'''
# prediction section below (uncomment this section to run the prediction for the next 30
days):
t = np.array(list(range(20, 50)))
plt.plot(t, trig_func(t, params[0], params[1], params[2], params[3]),
       label='Predicted trend')
plt.xlabel('Days elapsed since 21-10-2018') # labels x-axis
plt.ylabel('Apparent magnitude') # labels y-axis
'''

plt.legend(loc='best')
plt.show()
```

Bibliography

Total Internal Reflection, hyperphysics.phy-astr.gsu.edu/hbase/thermo/stefan.html. Acc. 18th December., 2017

"Absolute Magnitude | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/A/Absolute Magnitude. Acc. 12th January., 2018

"Blackbody Radiation | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/b/blackbody radiation. Acc. 19th December., 2017.

"Cepheid Variable Stars." *Density Definition*, astro.unl.edu/naap/distance/cepheids.html. Acc. 29th September., 2018

"Explore Math with Desmos." *Desmos Graphing Calculator*, www.desmos.com/. Acc. 3rd October., 2018

"Harvard Spectral Classification | COSMOS." *Centre for Astrophysics and Supercomputing*, astronomy.swin.edu.au/cosmos/H/Harvard Spectral Classification. Acc. 12th January., 2018

"Installation¶." *Matplotlib: Python Plotting - Matplotlib 2.2.3 Documentation*, matplotlib.org/. Acc. 21st January., 2019

"Inverse Square Law, General." *Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/Forces/isq.html. Acc. 19th December., 2017

Kenton, Will. "Overfitting." *Investopedia*, Investopedia, 13 Dec. 2018, www.investopedia.com/terms/o/overfitting.asp. Acc. 5th September., 2018

"Lecture 13: Binary Star Systems, Masses of Stars." *Culture, Creating Change, Physical Activity for All | Alberta Centre for Active Living*, sites.ualberta.ca/~pogosyan/teaching/ASTRO_122/lect13/lecture13.html. Acc. 4th March., 2018

"Luminosity and How Far Away Things Are." *European Imperialism*, pages.uoregon.edu/soper/Light/luminosity.html. Acc. 3rd February., 2018

"Mahasim (Theta Aurigae, 37 Aurigae)." *Universe Guide*, www.universeguide.com/star/thetaaurigae. Acc. 24th February., 2018

"NS Del." *SIMBAD Astronomical Database - CDS (Strasbourg)*, simbad.u-strasbg.fr/simbad/sim-basic?Ident=NS Del. Acc. 28th December., 2018

"NumPy¶." *NumPy - NumPy*, www.numpy.org/. Acc. 21st January., 2019

Revolvy, LLC. "'Alpha2 Canum Venaticorum Variable' on Revolvy.com." *Trivia Quizzes*, www.revolvy.com/page/Alpha2-Canum-Venaticorum-variable. Acc. 8th October., 2018

Roberts, Donna, and Frederick Roberts. *Rotation - MathBitsNotebook(Geo - CCSS Math)*, mathbitsnotebook.com/Algebra2/Statistics/STregression2.html. Acc. 10th August., 2018

"Scikit-Learn." *1.4. Support Vector Machines - Scikit-Learn 0.19.2 Documentation*, scikit-learn.org/stable/index.html. Acc. 21st January., 2019

"SciPy.org¶." *SciPy.org - SciPy.org*, www.scipy.org/. Acc. 21st January., 2019

"Slooh." *Slooh*, www.slooh.com/. Acc. 17th February., 2018

"Star Temperatures." *Total Internal Reflection*, hyperphysics.phy-astr.gsu.edu/hbase/wien.html. Acc. 18th December., 2017

"SuperWASP Observations of Variable Stars." *The Life of a Star*, www.astro.keele.ac.uk/workx/superwasp-variable-stars/Rotating.html. Acc. 27th June., 2018

"Theta Aurigae." *SIMBAD Astronomical Database - CDS (Strasbourg)*, simbad.u-strasbg.fr/simbad/sim-id?Ident=theta aurigae&NbIdent=1&Radius=2&Radius.unit=arcmin&submit=submit id. Acc. 21st February., 2018

"What Is Linear Regression?" *Statistics Solutions*, www.statisticssolutions.com/what-is-linear-regression/. Acc. 10th August., 2018

"What Is Underfitting | DataRobot Artificial Intelligence Wiki." *DataRobot*, 13 Dec. 2018, www.datarobot.com/wiki/underfitting. Acc. 6th September., 2018