Predicting the Habitability of Planets: A Comparison of Analytical and Machine Learning Models

Jagat Kafle

jagat.kafle_ug23@ashoka.edu.in

Nitin Jha

nitin.jha_ug23@ashoka.edu.in

Spandan Pandya

spandan.pandya_ug23@ashoka.edu.in

August 26, 2021

Abstract

The Circumstellar Habitable Zone (CHZ) refers to the circular region revolving around a star where temperature conditions allow for a possibility of liquid water. The presence of liquid water is closely associated with the possibility of life forms as we know it. The PHL database has already created a list of potentially habitable planets outside of our solar system. This paper attempts to present three models: two analytical, and one machine learning model in order to predict whether a planet lies in the habitable zone or not. The analytical model uses the Stephan-Boltzmann law and the idea of stellar flux to calculate habitability whereas the machine learning model uses a logistic regression consisting of 12 different parameters. This paper concluded that the analytical models have an accuracy of 86% and 89% respectively while the machine learning model has an accuracy of 95%.

To,

Our parents - our strength

Remembering,

Shri. Sunil Kumar Jha - The most luminous star

1 Introduction

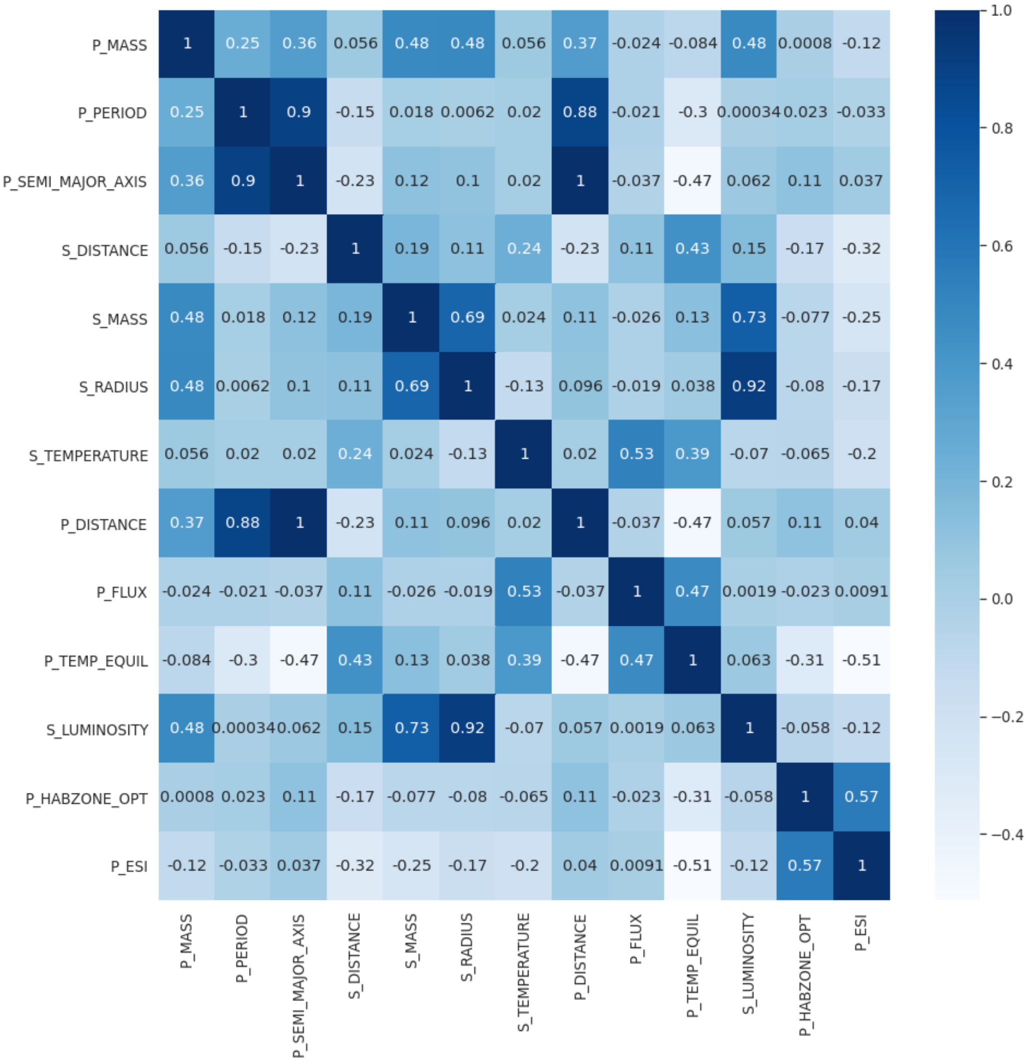
Planetary habitability is the measure of a planet's ability to maintain an environment necessary to sustain life. Based on the understanding of the earth's ability to sustain life, astrobiologists focus on some specific criteria to find the habitability of planets in the solar system and beyond. NASA's astrobiology roadmap [1] has defined variety of parameters that underpin habitability and life, including the presence and persistence of liquid water, potential free energy sources, physical and chemical environmental factors, and the presence of bioessential elements. The habitable zone (HZ) is a shell-shaped region of space surrounding a star in which a planet could maintain liquid water on its surface. Stellar characters like mass, luminosity and planetary characters like mass, radius, orbit and rotation are important in exploring the habitability of planets. Usually, earth like planets are the primary focus of exploration as they are highly likely to sustain life.

As of present, we have not yet discovered planets which are currently sustaining life. However, based on the physical environment, the availability of water, and other planetary and stellar features like mass, radius, luminosity - several planets have been classified as potentially habitable. The PHL exoplanet data-set [2] consists of around 4000 planets and lists these features along with several others, and classifies them as potentially habitable or not. This paper uses this data set in this paper, and built two analytical and one machine learning model to compare the accuracy of the classification results they yield.

For the two traditional (analytical) models, we have specified a habitable zone based on luminosity, albedo, effective temperature and temperature rise due to greenhouse effects. A planet which lies in this habitable zone means it's potentially habitable. In the machine learning model, the data set is filtered to remove null values and a logistic regression model is used to predict the habitability based on the correlation between the features of the planets and the stars. Results from the three models then are compared with the actual habitability status defined in the PHL catalog dataset to determine the accuracy of each model.

2 Data Visualization

As the PHL catalog consists of several null values, we removed these null values and the rows and columns associated with them to produce a final data set we could use in our models (See Appendix A). To see the relationship between several features like luminosity, radius, earth similarity index, a heatmap was constructed using python's seaborn library.



	P_MASS	P_PERIOD	P_SEMI_MAJOR_AXIS	S_DISTANCE	S_MASS	S_RADIUS	S_TEMPERATURE	P_DISTANCE	P_FLUX	P_TEMP_EQUIL	S_LUMINOSITY	P_HABZONE_OPT	P_ESI
P_MASS	1	0.25	0.36	0.056	0.48	0.48	0.056	0.37	-0.024	-0.084	0.48	0.0008	-0.12
P_PERIOD	0.25	1	0.9	-0.15	0.018	0.0062	0.02	0.88	-0.021	-0.3	0.00034	0.023	-0.033
P_SEMI_MAJOR_AXIS	0.36	0.9	1	-0.23	0.12	0.1	0.02	1	-0.037	-0.47	0.062	0.11	0.037
S_DISTANCE	0.056	-0.15	-0.23	1	0.19	0.11	0.24	-0.23	0.11	0.43	0.15	-0.17	-0.32
S_MASS	0.48	0.018	0.12	0.19	1	0.69	0.024	0.11	-0.026	0.13	0.73	-0.077	-0.25
S_RADIUS	0.48	0.0062	0.1	0.11	0.69	1	-0.13	0.096	-0.019	0.038	0.92	-0.08	-0.17
S_TEMPERATURE	0.056	0.02	0.02	0.24	0.024	-0.13	1	0.02	0.53	0.39	-0.07	-0.065	-0.2
P_DISTANCE	0.37	0.88	1	-0.23	0.11	0.096	0.02	1	-0.037	-0.47	0.057	0.11	0.04
P_FLUX	-0.024	-0.021	-0.037	0.11	-0.026	-0.019	0.53	-0.037	1	0.47	0.0019	-0.023	0.0091
P_TEMP_EQUIL	-0.084	-0.3	-0.47	0.43	0.13	0.038	0.39	-0.47	0.47	1	0.063	-0.31	-0.51
S_LUMINOSITY	0.48	0.00034	0.062	0.15	0.73	0.92	-0.07	0.057	0.0019	0.063	1	-0.058	-0.12
P_HABZONE_OPT	0.0008	0.023	0.11	-0.17	-0.077	-0.08	-0.065	0.11	-0.023	-0.31	-0.058	1	0.57
P_ESI	-0.12	-0.033	0.037	-0.32	-0.25	-0.17	-0.2	0.04	0.0091	-0.51	-0.12	0.57	1

Figure 1: The darker shades in the heat map indicate a stronger correlation. Each factor was chosen as a parameter for the machine learning model.

Visualising certain factors such as stellar mass, equilibrium temperature, earth similarity index, and Stellar luminosity in conjunction with the optimistic habitable zone given by the data allows for the emergence of visual trends.

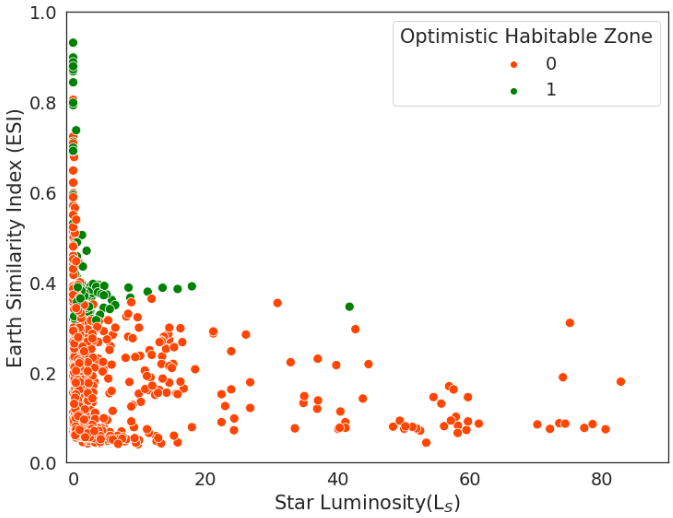


Figure 2: The green markers represent habitable planets and the orange markers represent non-habitable planets. This visualization implies that stars with higher luminosities are less likely to have planets that can be habitable. With stars having luminosity comparable to that of the sun, the ESI is close to that of earth's.

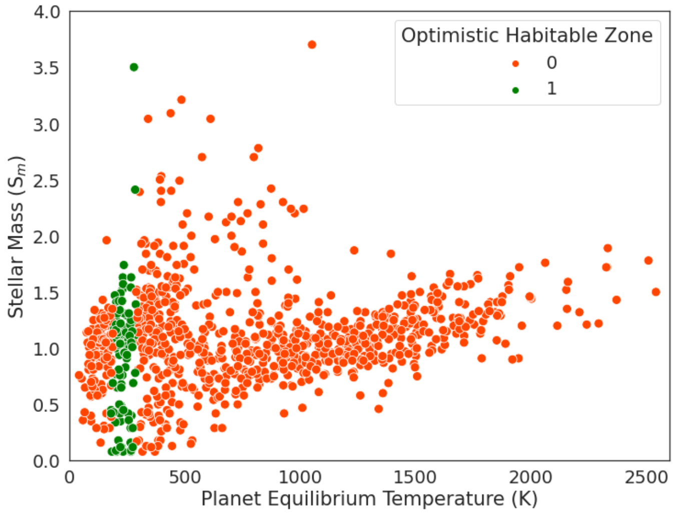


Figure 3: The green markers represent habitable planets and the orange markers show non-habitable planets. This visualization shows that planets with equilibrium temperature comparable to that of earth (273K−373K) are more likely to be habitable. Most of the data for habitable planets fall under 1.5 stellar masses, comparable to the sun.

3 Analytical Model 1 using the Stephan-Boltzmann equation

The existence of liquid water is the foremost criterion for the possibility of extraterrestrial life. Liquid water can only exist on a planet if its equilibrium temperature at a given pressure is between 273 degrees and 373 degrees Kelvin. The equilibrium temperature of a planet depends mainly on three parameters- Incoming stellar flux (Luminosity of the host star), albedo, and the heating due to the greenhouse effect. These parameters can be clubbed together in an equation. A planet that is at a distance d receives incoming radiation b depending on the total power (Luminosity) radiated by the star L.

$$b = \frac{L}{4\pi d^2} \tag{1}$$

Planetary features and atmosphere reflect a portion of this incoming radiation, this is indicated by the albedo effect A. The effective intensity that reaches the surface of the planet is given by:

$$b_{effective} = \frac{(1-A)L}{4\pi d^2} \tag{2}$$

The total energy absorbed by the planet (from the daylight side) should be equal to the total energy it would radiate as a spherical blackbody. Applying the Stephan-Boltzmann law gets the following results for the effective temperature of the planet:

$$\begin{aligned} P_{absorbed} &= P_{radiated} \\ \implies b_{effective} \times (\pi R^2) &= \sigma(4\pi R^2)T^4_{planet} \\ \therefore T^4_{planet} &= \frac{b_{eff.}}{4\sigma} \end{aligned} \tag{3}$$

The function for distance is:

$$d = \sqrt{\frac{L(1-A)}{4\pi \times 4\sigma T^4_{planet}}} \tag{4}$$

The input luminosity for the Python program (see Appendix A) used Luminosity in terms of solar luminosity, and gave the output distance in Astronomical Units (AU).

In order for the planet to be habitable, the equilibrium temperature should at least be 273 Kelvin, and at most be 373 Kelvin. Therefore the output for Habitability- $H(L)$ takes the form:

$$H(L_{star}) = \left[\sqrt{\frac{L(1-A)}{4\pi \times 4\sigma(273)^4}} \qquad \sqrt{\frac{L(1-A)}{4\pi \times 4\sigma(373)^4}}\right] AU$$

Another parameter that influences the equilibrium temperature of a planet is the existence of the greenhouse effect. This can be introduced in the equation as a parameter w. If the temperature change due to the greenhouse effect is w Kelvin, the heating of the planet due to blackbody radiation needs to be $(273 - w)$ K. For instance, if the earth naturally heats up around 32 degrees because of the greenhouse effect [3], even in areas with temperatures less than 273 K, such as 241 K would consist of water, since the additional heating due to the greenhouse effect would melt the ice. Therefore, adding the term w K modifies the model as follows:

$$H(L_{star}) = \left[\sqrt{\frac{L(1-A)}{4\pi \times 4\sigma(273-w)^4}} \qquad \sqrt{\frac{L(1-A)}{4\pi \times 4\sigma(373-w)^4}}\right] AU$$

The first analytical model assumed that in order for a planet to be habitable, it must show earth like values of Albedo and the greenhouse effect. Therefore, the first analytical model utilised $A = 0.28$ [4] and $w = 32$ K.

$$H(L_{star}) = \left[\sqrt{\frac{L(1-0.28)}{4\pi \times 4\sigma(273-32)^4}} \qquad \sqrt{\frac{L(1-0.28)}{4\pi \times 4\sigma(373-32)^4}}\right] AU$$

Applying this model to the data (see Appendix A) for 1001 different planets in order to predict whether they are habitable or not yielded the following confusion matrix.

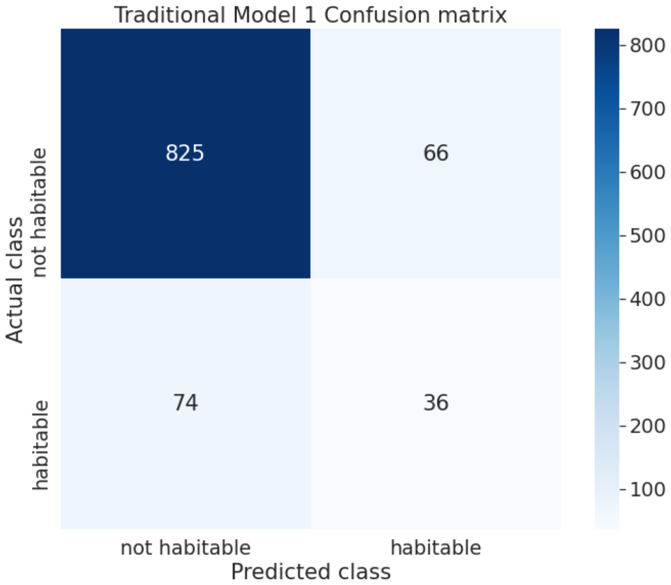


Figure 4: Analytical Model 1 confusion matrix

Inferring from the Confusion matrix, the model has an accuracy of 86% when compared to the data. However, out of the 110 actually habitable planets, only 36 were correctly predicted. This suggests that the data has a very high false negative rate. The model is rather simplistic, but still manages to correctly predict a large chunk of non habitable planets. The limitation of this model lies in the fact that only earth like values for albedo and the greenhouse effect were taken into consideration. Another limitation of this model is also that it assumes the planet itself to be a perfect blackbody, which may be an oversimplification of the situation.

4 Analytical Model 2 - A Modified Version of Model 1

The primary concern with the first analytical model was that it arbitrarily assumed values for albedo and the greenhouse effect parameters according to the earth. However in reality, planets come in diverse shape and form and are not bound by fixed values of albedo and heating due to the greenhouse effect. Therefore, the python algorithm used in the first model was modified in order to calculate habitable zones using a range of albedo values and a range of greenhouse effect values. The mathematical model remained the same, but the algorithm calculated different values based on changing parameters, and chose the combination that yields the lowest and the highest values for the habitable range. i.e., The computer program first calculates the data for all possible values of the greenhouse effect, keeping the albedo constant. Then the program repeats this step for a different set of albedo. The following ranges were chosen to calculate the Habitable zone:

$$w = [0, 100]$$
$$A = [0, 1)$$

Incorporating this change in the model led to the following Confusion matrix:

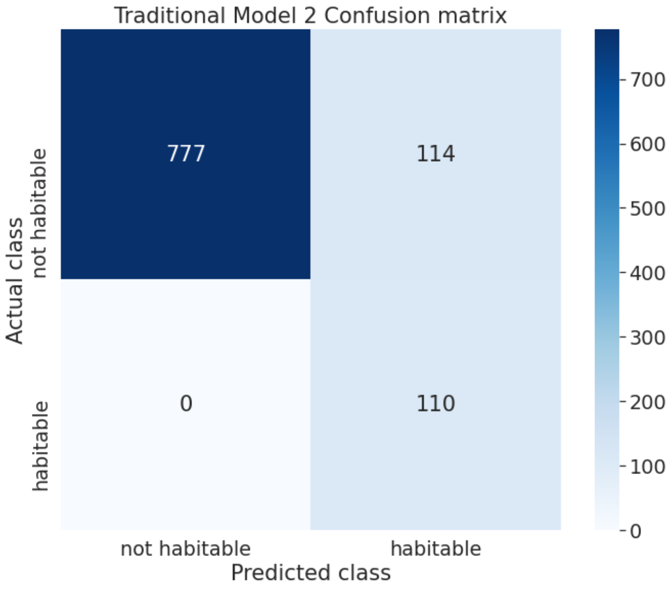


Figure 5: Analytical Model 2 confusion matrix

Comparing this model to the actual data suggests that this model is 89% accurate. This is a 3% improvement from the previous analytical model. In addition to this, the model resolved the issue of false negatives. This model successfully predicted all habitable planets with 100% accuracy. However, resolving the false negatives meant that the habitable range had to be widened. This has led to a significant increase in the false positives. One limitation of this model is that it only accounts for a greenhouse effect change of 100 K. While having a range for albedo and greenhouse effect significantly increases the accuracy, it still doesn't take into account the actual albedo of the planets. However, 89% accuracy rate is still remarkable for this type of a simplified model.

5 Machine Learning - Logistic Regression Model

The traditional models used the parameters and equations defined by us to find the probability of a planet being habitable or not, whereas our machine learning program decides on the parameters based on their correlation to the target variable. The model was trained for 70% of the data and the prediction accuracy was calculated for the rest 30% of test data. The classification was based on a binary logistic regression model, imported from the python library Sklearn. This model can be understood in two steps:

5.1 The Linear Combination of Parameters

All the input parameters were linearly combined with the biases associated with each of them and a binary Bernoulli response variable, Y was calculated.

$$Y = b_0 + b_1X_1 + b_2X_2 + ...b_nX_n$$

Here, b_0 can be interpreted as the log-odds of the event to be $Y = 1$ when all the predictors are zeros and b_n shows the increase in the log-odds of $Y = 1$ when the predictor, X_n is increased by this factor b_n. The input parameters are the 12 parameters such as S_MASS, S_LUMINOSITY, P_PERIOD, P_DISTANCE, etc. These are combined to give the response variable which is then used to calculate the value of this response variable, Y.

5.2 Passing through a Sigmoid function

Now, this Y is being passed through a special function known as the Sigmoid function. This step is then used to calculate the binary output for the code. In our model, the outputs were, 0 beings 'not habitable' and 1 denoting 'habitable'.

This output, O, can be represented by the equation:

$$O = \frac{1}{1 + a^{-(b_0+b_1X_1+b_2X_2+...b_nX_n)}}$$

This equation, thus, taking into account the response variable predicts the binary output as 0 and 1. The above equation can also be written as:

$$O = S_a(b_0 + b_1X_1 + b_2X_2 + ...b_nX_n)$$

O can give the following outputs: 0 and 1. The input columns (the parameters discussed above) were passed as X_n's for particular planets, and the value of O was determined to be either 0 or 1. The model was trained for 70% of data for 200 iterations and then the output was found for the remaining 30% data (see appendix). This set of outputs for all the planets used were found and compared to the existing data about their probability of being habitable. The factors such as b_n, a are automatically adjusted by the regression model according to the input parameters. A confusion matrix was constructed for comparing the accuracy of the Machine Learning model with the PHL data.

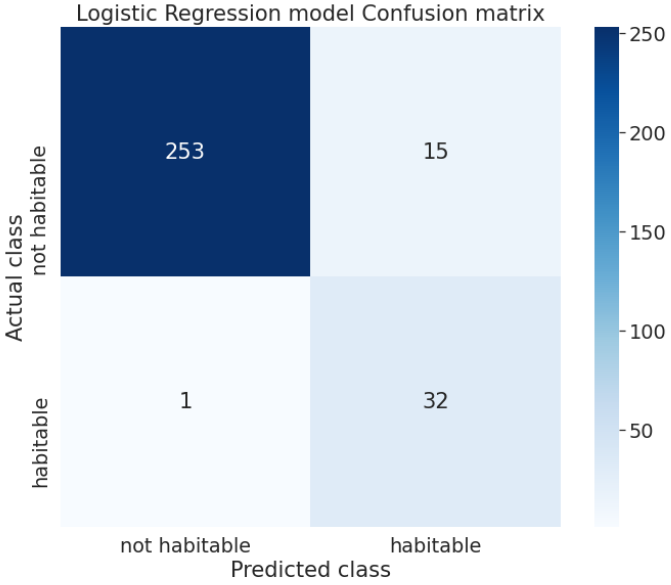


Figure 6: Machine Learning Model confusion matrix

When compared to the predetermined data, this program accurately predicts about 95% of the planets to be habitable or not. This shows a 10% and 7% increase in accuracy compared to the traditional model 1 and 2 respectively. The final data determined by the machine learning suggests a True-negative rate to be 97%, and the True-positive rate to be 94%.

6 Conclusion

The paper concludes that the use of Machine Learning in order to predict whether a planet is potentially habitable is the most accurate, with an accuracy of 95% The analytical models discussed in the paper have the accuracy of 86% and 89% respectively. The reduced accuracy in analytical models stems from the fact that these models do not utilise specific data relating to planets, but uses arbitrary or earth-like values. The analytical model 2 that utilises a range of albedo values and greenhouse values is an interesting model because in spite of simplistic assumptions, the model has no false negatives and has an accuracy of 89%. These analytical models and machine learning models can be made more accurate by the incorporation of atmospheric modelling through the data of different gases. The use of these three models use college level physics and programming languages, making them ideal for initial astronomy explorations. The models can also be used for segregating large chunks of planetary data into more usable sets for specific habitability based calculations.

7 Acknowledgements

The authors are extremely grateful to Prof. Dr. Vikram Vyas for his prompt feedback and insightful suggestions. His guidance has been invaluable, the authors express their heartfelt respects.

References

[1] "NASA Astrobiology." *astrobiology.nasa.gov*, 2021, `https://astrobiology.nasa.gov/about/astrobiology-strategy/`.

[2] "PHL's Exoplanets Catalog - Planetary Habitability Laboratory @ UPR Arecibo." *phl.upr.edu*, 2021, `http://phl.upr.edu/projects/habitable-exoplanets-catalog/data/database`.

[3] "What is the greenhouse effect?" *climate.nasa.gov*, 2021, `https://climate.nasa.gov/faq/19/what-is-the-greenhouse-effect/`

[4] Goode, P.R.; et al. "Earthshine Observations of the Earth's Reflectance". *Geophysical Research Letters*, 2001, 28 (9): 1671–1674. `Bibcode: 2001GeoRL..28.1671G`. `doi:10.1029/2000GL012580`.

A Code Repository

Please refer to the github repository link below to access the code used for this paper. `https://github.com/jakafle/sumproj-habitable`