Astron 98 Final Project: Analyzing the Distance from Earth vs Earth Similarity Index (ESI) of Potentially Habitable Exoplanets

Lauren Sandberg and Grace Jost

1. Introduction:

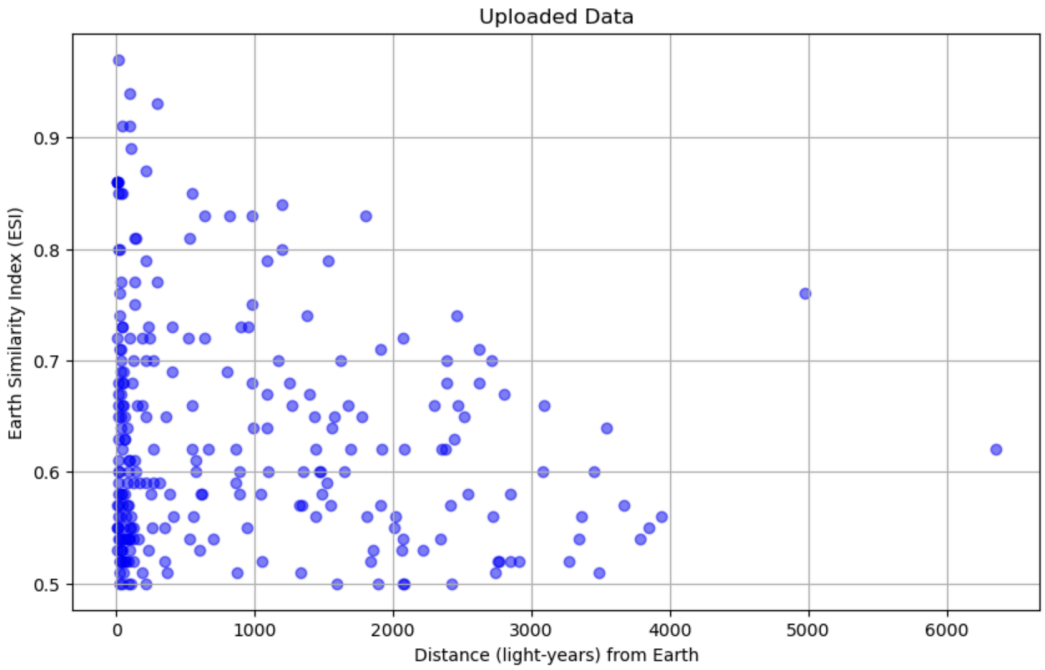
In our project proposal, we will outline the approach we will take to analyze the distance from earth compared to the ESI of potentially habitable planets in our galaxy. We want to determine if there is any correlation between the ESI and the distance from Earth for each of the potentially habitable planets in the data set. This will include data analysis, data generation, applying quality filters, fitting data with respect to error, and explaining the model fit.

2. Chosen Phenomenon and Data Source:

The chosen phenomenon for this project is the distribution of plausibly habitable planets in the Milky Way galaxy, aside from our own planet. To study this, we will use data from the Habitable Worlds Catalogue. From this catalog, there are 257 data points of potentially habitable planets (ESI(earth similarity index)>0.5)

3. Equation to Fit Data

Because of the way our data was natured, there is no way that we would “fit” the data. It looks extremely randomized until you look at the last part of the graph, where an interesting drop in data points happens. Because of this, we decided to plot an upper limit of the graph to see what distance compares to what maximum ESI, after properly removing outliers. For this upper limit we will use a simple line of best fit y=mx+b equation in the form slope times the distances of all the points plus the average y-intercept. Down below, we have inserted a picture of our raw data plotted.



4. Data Generation for Testing:

Random data was not needed to be generated because of the nature of our problem. Our plotted data shows that to fit data, we do not want to do a fitted correlation, but it's an upper limit so no random data.

5. Data Filtering

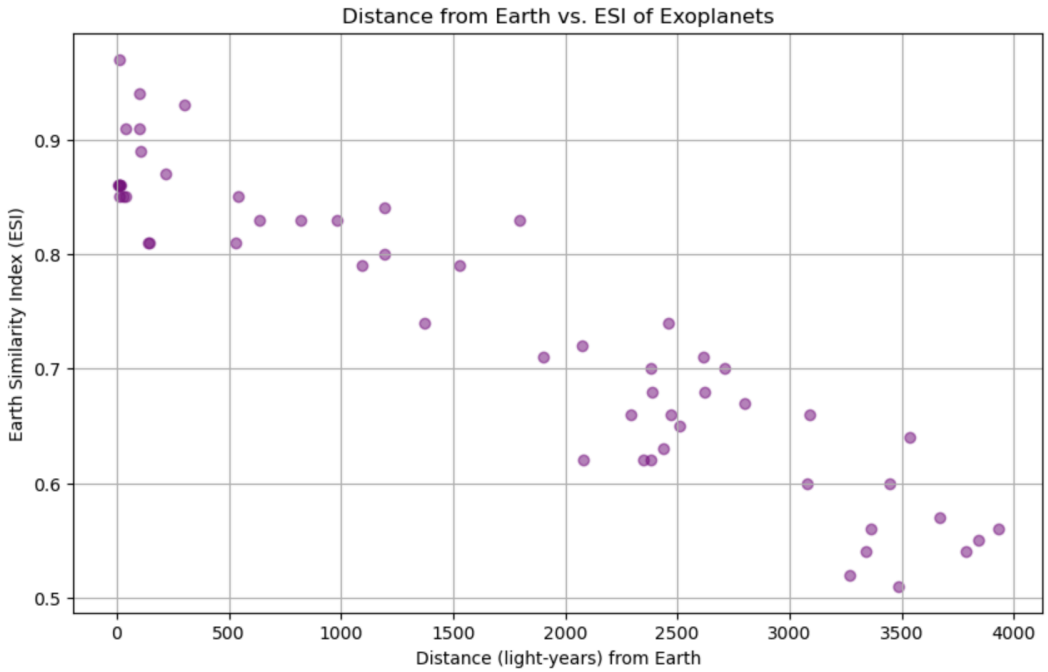
To ensure the quality and reliability of the dataset, we applied several quality filters and data pre-processing steps:

Outlier Removal:

- According to the website we used, the habitable worlds catalog, potentially habitable planets are those with an Earth similarity index of 0.5 or greater. The data set we found has data with ESIs lower than 0.5. So, we will neglect this data in order to only be looking at potentially habitable planets. We took this data out before even plotting our data because we knew we didn't want to include any planets that weren't 'certifiably potentially habitable'.
- After plotting, we noticed that some distances were extreme as well and could be considered outliers. We ended up filtering out any distance above 4000 light years.
- After plotting, we also decided that we wanted to do an upper limit line of best-fit, so data that would not pertain to an upper limit also needed to be deleted. We ended up

filtering out data in large blocks. We removed from when x is less than 2000 and y is between 0.5 and 0.7, when x is less than 1000 and y is between 0.7 and 0.8, and lastly, when x is between 2000 and 3000 and y is between 0.5 and 0.6.

Here is a graph of our final- filtered - data:



6. Data Fitting with Error

After taking out the outliers and removing points so that only the data points that have to do with the upper-limit remain, all we had to do was create a line of best fit to these leftover data points.

7. Explanation of Model Fit

Since there is no apparent correlation between distance from Earth and ESI of exoplanets, we used a line of best fit on our filtered data. This serves to summarize the general trend or relationship between our two variables (distance and ESI) in a simple and interpretable way. It helps us understand visually if there's any correlation. The line of best fit is a straight line that minimizes the total distance between the line and all the data points. In our graph, it represents the average trend of how ESI changes with distance.

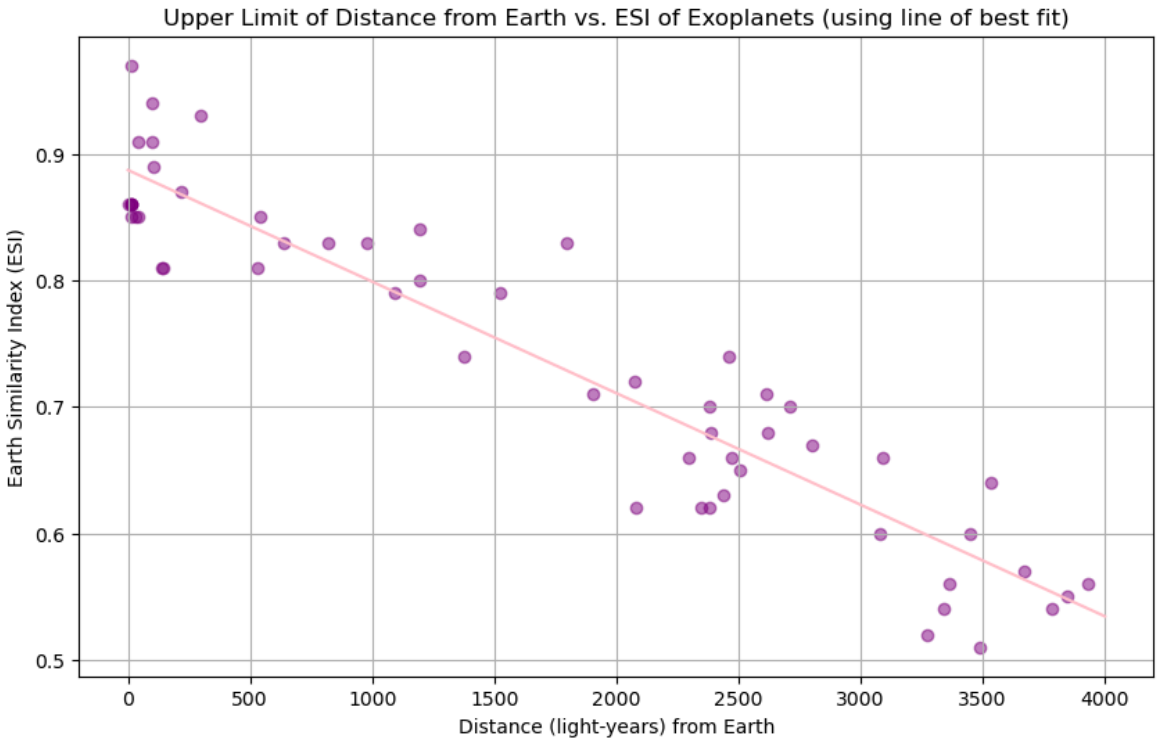
The slope and y-intercept of our best fit line are:

Slope: -8.819684228523461e-05 = -0.00008819684228523461
Y-intercept: 0.8871319925498681

These give us this equation:
Y ≈ -0.000088x + 0.89

Since our slope is very close to 0 (although slightly negative which means a negative correlation), it suggests slight correlation between the upper limits of our two variables, whereas distance from Earth increases, the maximum ESI decreases.

Final filtered data with line of best fit:



8. Conclusion:

The line of best fit used as an upper limit provides us with a useful tool to estimate for what distances from earth, the maximum ESI is. Overall however, there is no direct relation of the distances of these potentially habitable exoplanets from earth compared to their exact ESIs, there is only the upper limit. Additionally, with our data, we noticed there was a large

concentration of data points closer to Earth. We determined that this wasn't because there are more exoplanets close to us, but solely because they are easier to 'discover' when they are closer in distance to us.

Throughout our project we had a couple of hiccups down the road. Upon first look at our data, we had zero clue how to fit it in the sense that we weren't sure which way would be best with the appearance of lack of correlation. We met with course administrators and showed them our graphed data points, to which they suggested the upper limit line. After this help, it was smooth sailing.

Overall there was a lot of trouble shooting with code and we learned a lot about solving these problems and overtime, it got easier and easier to figure out how to locate our mistakes in code and rework things.

This project challenged us, yet helped both of us understand plotting, cleaning, and fitting data on a higher level.

Though we did not find any correlation in our data, this is an interesting observation on its own, and is an important lesson because this could very well happen in other astrophysics research projects.