Finding Stars with Earth-like Planets

Classifying Stars in the TESS Input Catalog According to Their Likelihood of Harboring Planets Similar to Earth

Peter Peirce
Masters Student, Courant Institute
New York, New York
pp1844@nyu.edu

***Abstract*—This paper describes the results of a large-scale data mining effort with the goal of identifying those stars with properties favorable for harboring Earth-like planets. The paper describes the efforts that went in to data collection, reduction, clustering, and classification of the NASA Exoplanet Archive and the TESS Input Catalog. After studying a 20° band of the sky, 388 star candidates were determined with moderate certainty to harbor planets of interest. The primary conclusion drawn by this study is that statistical analysis of exoplanet data can provide a valuable starting point for future sky surveys.**

I. Introduction

Prior to 1992 there were no confirmed observations of any planets outside of our solar system. Today there are nearly four thousand known exoplanets. By all estimations there are more planets in the universe than there are stars. Many of those planets look a lot like Earth, and with those similarities come the chance that they harbor life.

In order to more carefully study these planets beyond our solar system, scientists have been sending increasingly powerful telescopes into orbit. The most successful of these surveying telescopes so far has been Kepler. It discovered twenty-six thousand exoplanets in its nearly ten years of operation. Kepler's mission ended this year, but in its wake even more powerful telescopes are being deployed. The Transiting Exoplanet Survey Satellite (TESS) was launched this year. It started operations in the summer and over the next two years it is expected to discover more than twenty thousand new exoplanets within three hundred light years from Earth. In order to do this, the scientists behind the mission have compiled a candidate list of stars that may be of interest to study.

Of course, Earth-like planets won't be seen orbiting all of these stars. Having a method to determine which stars have a greater probability of harboring detectible Earth-like planets would allow researchers to focus their efforts on those stars when profiling TESS data. Not only would this result in the confirmation of more Earth-like planets, giving the scientific community a more robust understanding of these planets, but it would provide future missions a deeper catalog of Earth-like planets to select from, resulting in cascading scientific benefit for years to come.

II. Related Work

A. *Data Mining for Extra-Solar Planets [4]*

The authors of this paper, written in 2007, performed a data mining project with the goal of finding favorable stars for future SETI-like searches and extra-solar planet searches. They mined the USNO-B catalog of more than one billion astronomical objects to generate a reduced best-candidate list of a few hundred stars. As of the writing of their paper, the authors felt that the existing exoplanet catalogs were biased towards large planets with short orbital periods and were interested in discovering stars that were likely to harbor small, terrestrial planets orbiting their stars at moderate distances.

They proceeded to perform K Means clustering on the USNO-B catalog across a four-dimensional parameter space which encompassed three parameters that described the stars light intensity in specific color bands, a surrogate for describing the stars age and sequence type) and one parameter that described its proper motion across the sky (a surrogate for the stars distance from Earth). After performing K Means, three contiguous clusters (and one noise cluster) were discovered.

They then mapped 129 stars from the extrasolar catalog onto the four dimensional parameter space and found that these stars mapped almost exclusively to cluster 2, which contained mid-range, solar-like B2-R2 colors and moderate-to-high proper motion, which suggests the stars are fairly close to Earth.

B. *The stellar-metallicity—giant planet connection [2]*

The authors of this paper, written in 1996, were among the first to write in detail about the connection between a stars metallicity and the planet formation process. In this paper, they analyzed the spectroscopic abundance of stars υ And and τ Boo, both considered metal rich with [Fe/H] measurements above 0.20. They suggested that the high metallicity of these stars is the result of self-pollution during the planet formation period, not a result of the composition of the original interstellar cloud that preceded the planet formation process.

C. *A Quantitative Comparison of Exoplanet Catalogs [1]*

This paper, written in 2018, compares the four most commonly used exoplanet catalogs. Good agreement was found across planetary parameters and stellar properties, with minor differences that the authors concluded would be unlikely

to significantly affect any statistical studies that used any of these catalogs.

III. CRISP-DM PHASES

A. Problem Understanding

There are billions of stars in our galaxy, each with the potential of harboring Earth-like planets. Exoplanet hunting telescopes cost billions of dollars to construct and launch into orbit and have limited useful mission lifetimes. Every moment one of these telescopes is looking at a star without an observable planet it is not only wasting thousands of dollars, but also losing time that could be spent observing stars with planets and gathering more data from them. Any method to reduce this time spent looking at uninteresting stars would be immensely valuable both financially and scientifically.

The object of this project is to dramatically reduce the number of stars that exoplanet hunting satellites have to observe in order for them to detect interesting Earth-like planets. While no statistical analysis will perfectly predict stars of interest, a prediction with some degree of confidence would provide a good starting point, allowing satellites to prioritize those stars thought to harbor Earth-like planets and study those stars first and for longer periods of time. This could potentially result in a dramatically improved return on investment for these expensive satellite observational missions.

Success for this project is be defined as creating an analytical model that could predict, with a high degree of confidence, which stars harbor Earth-like planets. While statistical tests can be performed on the data describing already confirmed exoplanets to determine the quality of the model as measured by precision and recall, the real test will be performed as TESS releases the results of its survey over the next two years and the planets that it finds are compared to the predictions made by the model. If the model is found to be reliably predictive, it may be employed by future missions to cull the list of stars being studied, allowing more time to be spent studying more valuable stars.

B. Data Understanding

For this project, two datasets were used across two different stages. In the first stage, known exoplanets were clustered into two groups and the stellar parameters of the host stars of those planets were extracted to be used in the following stage. This exoplanet data was obtained from the NASA Exoplanet Catalog (NEP), a publicly available dataset containing 3826 planets at the time of download. This catalog is maintained and used by the American space agency alongside CalTech. Each planet was described by 142 parameters, including stellar parameters describing the host star of the planet. For the purposes of clustering, only those planets with both planetary mass and planetary radius were of interest, as these are the properties that can be used to describe the planet in terms of its similarity to Earth (small and rocky) and Jupiter (large and gaseous), the two prototypical planets of our solar system. Planets missing either of these parameters could be discarded. 3022 planets in the database had a radius measurement, and 1416 planets had a mass measurement. There were 624 planets with both parameters. The radius, measured in units of Jupiter radius, had a mean value of 0.36 J, a min of 0.03 J, and a max of 6.9 J. The mass, measured in units of Jupiter mass, had an average of 2.6 J, a min of 0.00006 J, and a max of 55.6 J.

The stellar properties were used in the classification step. Of interest were distance from Earth measured in parsecs, surface temperature measured in Kelvin, mass measured in units of Solar mass, radius measured in units of Solar radius, log of surface gravity, luminosity, and metallicity as the ratio of Iron to Hydrogen in the star.

Looking at the NASA Exoplanet Catalog, 3813 entries had distance data with a min of 1.29 pcs, a max of 8500 pcs, and an average of 623 pcs. 3670 entries had temperature data, with a min of 575 K, a max of 57,000 K, and an average of 5495 K. 3052 entries had stellar mass data, with a min value of 0.02 sol, a max of 23.56 sol, and an average of 1.02 sol. 3568 entires had stellar radius data, with a min value of 0.04 sol, a max value of 71.23 sol, and an average of 1.55 sol. 3441 entires had a surface gravity measurement, with a min value of 1.2 log(solar), a max of 5.52 log(solar), and an average of 4.36 log(solar). 534 entries had a luminosity measurement, with a

Table 1. Parameters of NASA Exoplanet Catalog Stellar Data

	Distance (pc)	*Temperature (K)*	*Mass (Sol)*	*Radius (Sol)*	*Gravity log(solar)*	*Luminosity log(solar)*	*Metallicity (dex)*
Min Value	1.29	575	0.02	0.04	1.2	-3.48	-0.89
Max Value	8500	57000	23.56	71.23	5.52	3.01	0.69
Average	623	5495	1.02	1.55	4.36	-0.059	0.016

Table 2. Parameters of NASA Exoplanet Catalog Stellar Data

	Distance (pc)	*Temperature (K)*	*Mass (Sol)*	*Radius (Sol)*	*Gravity log(solar)*	*Luminosity log(solar)*	*Metallicity (dex)*
Min Value	7.47	4000	0.49	0.38	0.034	-2.46	-2.46
Max Value	3027.9	7801	2.79	42.2	242.4	242.4	5.58
Average	375	5668	1.08	3.4	3.76	2.51	1.03

min of -3.48 log(solar), a max of 3.01 log(solar), and an average of -0.059 log(solar). Finally, 2758 entries had a metallicity measurement, with a min of -0.89 dex, a max of 0.69 dex, and an average of 0.016 dex.

For star data, the TESS Input Catalog (TIC) was used. This is the list of stellar candidates for the new TESS mission that began operation earlier this year. While not all stars on this list will be observed by TESS during its two year primary mission, TESS will only observe stars that are on this list. While the TIC contains data for all 180° of the sky, because of the large nature of this data — totaling over 70 GB — a 20° subset of the catalog, from -90° to -70°, was examined. This subset of the sky contained a total of 11,167,602 stars. After removing all stars missing any of the seven parameters that were used for classification, only 33,957 stars remained — a reduction of the data to 0.3% of its original size.

Looking at the same parameters in this subset of the TIC after data reduction as were characterized above for the NASA Exoplanet Catalog, the minimum distance was 7.47 pcs, a max of 3027.9, and an average of 375. The min temperature was 4000 K, the max was 7801 K, and the average was 5668 K. The min stellar mass was 0.49 sol, the max of 2.79 sol, and an average of 1.08 sol. The min stellar radius was 0.38 sol, the max was 42.2 sol, and the average was 3.4 sol. The min surface gravity was 0.034 log(solar), the max was 242.4 log(solar), and the average was 3.76 log(solar). The min luminosity was -2.46 log(solar), the max was 242.4 log(solar), and the average was 2.51 log(solar). The min metallicity was -2.46 dex, the max was 5.58 dex, and the average was 1.03 dex.

A summary of these findings are found in Tables 1 & 2 on the previous page.

C. *Data Preparation*

Planets from NEP were clustered across only planetary mass and radius. These two parameters were chosen because they are the most effective parameters that can be used to compare exoplanets to plants within our solar system. By describing the size of the planet, we can quickly deduce other properties, such as whether the planet is rocky or gaseous. Specifically, based on what we know from our own solar system, we know that smaller planets are rocky and large planets as gaseous. Because, as best we know, life is more likely to occur on rocky planets it makes sense to focus our research on these smaller planets. Larger planets, while easier to detect, are of less interest.

In the model, data was read in to memory from a .tsv file row-by-row and those planets without values for mass and radius were immediately discarded. Because both the mass and radius values have ranges that span many orders of magnitude, and because these two fields do not have comparable units of measurement, the values were standardized using the Z-Score (new value = (old value - mean) / standard deviation). This way, each value could be clustered according to its distance from the average value of the parameter, rather than as an absolute distance.

As previously mentioned, data was thrown out if not all parameters of interest were included. However, this was not the original method that was used to handle missing data. The first method that was tried was filling in missing values with the averages for that particular parameter across all rows. While this may have worked for some parameters that were only missing from a few rows, in the case of parameters such as luminosity only 534 of 3813 objects had data. Filling in 3279 values with the same average value would have resulted in severe distortion of the model. The vast majority of the data would be completely fabricated. After deciding that this was not a viable option, it was found that even if every row that did not have all parameters were removed there would still be hundreds of viable rows that could remain. It was then decided that that provided enough diversity in the data to carry on in this fashion.

Only 20° of the total 180° of TIC data was used for classification. This was done in part because even the 20° of sky data that was used contained more than eleven million stars. After removing all stars with missing data, more than thirty three thousand stars remained. As will be seen in the results, this provided a significantly large sample to predict hundreds of stars that may harbor Earth-like planets. Future research may wish include the full sky database or specific regions of the sky studied by future missions.

D. *Modeling*

There were two stages to the modeling of this project: clustering the NASA planetary data by planetary mass and radius into Earth-like and Jupiter-like clusters, and classifying TIC stellar data into those clusters using the stellar parameters of the NASA data. The goal was to discover stars in the TIC catalog likely to harbor Earth-like planets.

Clustering on the NASA catalog was carried out using K Means with k = 2. The decision to set k to 2 was simple: there

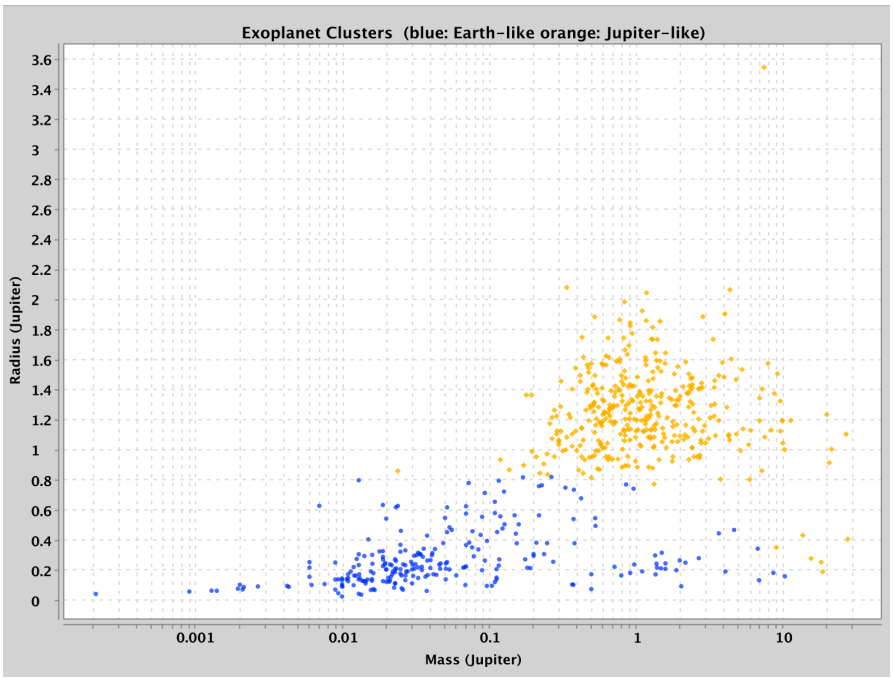

Table 3. Test Results from KNN Classification for Earth-like Prediction

	All Parameters	*- Distance*	*- Distance, Temperature*	*- Metallicity, Luminosity*
Recall	0.13	0.33	0.67	0.13
Precision	0.40	0.71	0.77	0.40
F1 Score	0.20	0.45	0.71	0.20

are two broad categories of planets, Earth-like and Jupiter-like. Initial attempts to cluster the planets did not provide usable results. However, after normalizing the data using Z-Scores, reasonable clusters that evenly divided the dataset into two subsets that appeared to follow expected Earth-like and Jupiter-like definitions were obtained. Normalization was useful in this case because the two parameters, mass and radius, had notably different scales, mass ranging from 0.0001 to 1, radius ranging from 0.1 to 1.

One downside to using K Means is noise points: planets that don't fit cleanly into either category of Earth-like or Jupiter-like, were forced into one of those two categories. An implementation of DBSCAN, which would address this issue, was written, but ultimately not used due to time constraints and incompatible output types. Future research may want to use DBSCAN to obtain tighter definitions of Earth-like and Jupiter-like planets. However, for the purposes of this investigation, the results obtained by K Means with k = 2 were sufficiently useful.

Once the clusters were determined, the next step was to classify the star data from the TESS Input Catalog according to the clusters determined in the last step using the stellar parameters of the planets in those clusters. KNN was the classifying algorithm that was implemented for this purpose. Different values of k, ranging from 2 to 10, were tested and ultimately no difference was found amongst them. As such, 3 was chosen as the value for k.

Testing the effectiveness of the classifier was done by clustering a large subset of the NASA Exoplanet Catalog data and putting aside the remaining data for classification. A confusion matrix of the expected clusters and the generated clusters was produced, as well as recall, precision, and f1 scores.

Early runs of the classifier used all seven stellar parameters, however through testing it was determined that stellar distance and temperature were not valuable to the model and were discarded. Running the testing procedure while including all parameters resulted in a F1 score of 0.20 for predicting stars with Earth-like planets. Removing just the distance parameter increased the F1 score 0.45, and removing both distance and stellar temperature increased the F1 score to 0.71. Attempts at removing other parameters did not significantly affect the F1 score.

D.*Evaluation*

While initial runs of the classifier which included all seven original stellar parameters were discouraging, later runs which removed temperature and distance parameters resulted in output that warrants cautious optimism. Recall, precision, and F1 scores all around 0.70, while not conclusive by any means, at the very least suggest that future research into this line of inquiry is necessary. A more comprehensive suite of algorithms (not written by hand in Java by an overworked graduate student) could be used to suss out more conclusive relationships between stars and their likelihood of harboring Earth-like planets. As already mentioned, DBSCAN may be a more appropriate clustering algorithm for this dataset, as the forced inclusion of noise by K Means may have negatively impacted the results to some degree. Experimenting with clustering algorithms designed for use with astronomical and exoplanet data [5] may be of interest as well.

Future exploration of this topic ought probably not to follow in my footsteps of implementing the algorithms by hand in Java. Use of R and standard statistical libraries is likely warranted.

E. *Deployment*

For now, deployment of this model into any real world application is very unlikely. However, in the future after refinement and peer review, the team managing TESS and the teams managing future exoplanet missions may wish to use the ideas presented here as a way of focusing their search efforts. Applying this model to their initial stellar candidate list of billions of stars may help them pick out a few hundred top candidates that they wish to spend more time investigating in great detail.

IV. CONCLUSION

The problem of data reduction is not new. And no field is less a stranger to it than astronomy. Relatively new to this field is the area of exoplanet research. As more and more satellites begin to search the skies for interesting planets orbiting other stars, a way to increase the efficiency of that search and allow astronomers to select a handful of top candidates likely to harbor Earth-like planets from lists of billions of stars is of increasing importance. This paper described the results of a model that was designed to do just that. By applying K Means

Table 4. Properties of Stars Predicted to Have Earth-like Planets

	Distance (pc)	*Temperature (K)*	*Mass (Sol)*	*Radius (Sol)*	*Gravity log(solar)*	*Luminosity log(solar)*	*Metallicity (dex)*
Average	258	5275	0.99	1.75	10.78	0.30	0.08

clustering and KNN classification to astronomical data, 664 stars were selected from millions as being the most likely to harbor Earth-like planets.

Future research down this line of inquiry ought to learn from my mistakes. More selective clustering algorithms should be used in order to more selectively define what an Earth-like planet is. And while feature selection did result in a 0.5 point improvement in F1 scores, more exploration into other stellar parameters may result in even better outcomes.

References

1. D. Bashi, R. Helled, and S. Zucker. A Quantitative Comparison of Exoplanet Catalogs. Geosciences 8 (2018), 325.
2. G. Gonzalez. The stellar metallicity-giant planet connection. Mon. Not. R. Astron. Soc. 285 (1997) 403-412.
3. G.R. Ricker et. al. The Transiting Exoplanet Survey Satellite. Proc. of SPIE 9904 (2016)
4. K.D. Borne, and A. Chang. Data Mining for Extra-Solar Planets. ASP Conference Series 376 (2007) 453-456.
5. W.L. Hung, S.J. Chang-Chien, and M.S. Yang. An intuitive clustering algorithm for spherical data with application to extrasolar planets. Journal of Applied Statistics 42 (2015) No. 10 2220-2232.