ESTIMATING THE NUMBER OF EARTH-SIZED HABITABLE PLANETS IN THE MILKY WAY GALAXY

A. Calhoun[1], I. Fowler[1], E. Lee[1], M. Nagarathnam[1], T. Ptak[1], and I. Son[1]

[1] *Illinois Mathematics and Science Academy, Aurora, IL 60506, USA.*

ABSTRACT

The estimated number of Earth-sized habitable planets in our galaxy was derived from the number of Earth-size habitable planets discovered by the Kepler space telescope and through an understanding of Kepler's detection efficiency of these planets. Although the number of Earth-sized planets detected by Kepler is already known, the telescope's instruments are not perfectly precise and cannot detect many planets, particularly those that are too small and/or at too great a distance for Kepler's detection capability [1]. Simulations of artificial planet transits were created and analyzed by a trained machine-learning algorithm to approximate the detection limits of the Kepler space telescope. The simulated transits were created using characteristics such as planetary and orbital radii within set chosen boundaries of habitability, while the machine-learning algorithm was trained on known null and positive stars with exoplanets from the NASA Exoplanet Archive. Through the combination of the Kepler detection efficiency, geometric factors, and an estimated number of stars within the galaxy, a scaling factor for the number of exoplanets confirmed through the first Kepler mission was found. From this, an estimate of Earth-sized habitable planets including those not detected by Kepler in the Milky Way was derived to be 2.7113145e+12, an obvious indication of experimental error.

Subject heading: methods: data analysis - planets: detection - techniques: photometric

1. INTRODUCTION

1.1 *Rationale*

The *Kepler Space Telescope* launched in May of 2009 and has since gathered photometric data on more than half a million stars and has led to the confirmation of more than 3000 exoplanets [2]. The photometric light curves produced by the *Kepler Space Telescope*, particularly in the early stages of data analysis, created a unique technical condition where both large crowdsourced data analysis and neural networks work together to investigate exoplanet candidates [3]. Deep convolutional neural networks created by NASA and Google have provided a model of effective light curve detection that can be scaled for future photometric detection research [4].

As a primary goal of the *Kepler Space Telescope* was to understand and quantify exoplanets as well as those that maybe within ranges of characteristics of habitability, it is then necessary that the vast amounts of collected is thoroughly analyzed. To this day, many of Kepler's exoplanet candidates are still unconfirmed [5].

1.2 *Background*

Data for this study came solely in the form of photometric light curve spreads. Broadly speaking, to collect photometric distributions and, potentially, observe planets, the stellar flux of a star is taken over a given period of time, often over a 30 day period [6]. When statistically

significant “dips” in stellar light flux are found, replicated, and analysed the existence of an exoplanet is confirmed [7]. Our analysis looked at many different characteristics of singular non-binary star systems from the “NASA Exoplanet Archive” in order to determine the habitability of the planet. These factors included luminosity planet's parent star, planetary radii, orbital radii, and potential atmospheric composition. To account for the detection errors and technological limitations of the Kepler space telescope, a detection scaling factor was needed to properly account for potential missing exoplanets. We determined the Kepler detection efficiency by training a machine learning algorithm using lightcurves from the Kepler database. This detection efficiency along with geometric factors, that account for the relative positioning of the telescope in relation to potential star systems with exoplanets, was used to estimate the number of Earth-sized habitable planets within the Milky-Way galaxy.

2. EQUATIONS AND DETERMINING HABITABILITY METHODS

2.1 *Baseline Equation*

The goal of this analysis is to understand both the Kepler detection efficiency and to, in turn, estimate the number of Earth-sized habitable planets P_{net}. To accomplish this, a general formula was conceived that would both account for the known constants of the geometric efficiency E_g, ratio of habitable planets to total stars seen by Kepler, P_h/S, and number of stars within the galaxy, S_{mw}, as well as for the enquired Kepler detection efficiency, E_k.

$$P_{net} = \frac{P_h}{S} \times S_{mw} \times \frac{1}{E_g} \times \frac{1}{E_k} \tag{1}$$

The given equation was used for the final estimation of habitable planets in the Milky Way galaxy and was the basis of our research. The provided terms will be discussed in greater detail in the following sections of this paper.

2.2 *Equations for Habitability*

Two primary restrictions were created in determining the ratio of “habitable”-confirmed planets- to the total number of stars analysed. The first condition was the exclusion of all planets whose radii exceeds that of 1.8 Earth Radii or is smaller than 0.5 Earth Radii [9]. The second condition was habitable orbital radii, or zones where the planet must exist to maintain a surface temperature suitable for life, in meters from the parent star. Based on the effective temperature of the main sequence parent star, our conditions established a planetary temperature range from 225K to 350K that excludes any potential planetary atmospheric effects. The inner r_i and outer r_o radii, measured in meters, of the habitable zones based on the effective temperature T_e, measured in Kelvin, are given below. To simplify our calculations, we took main sequence stars and fit their radius to their temperature.

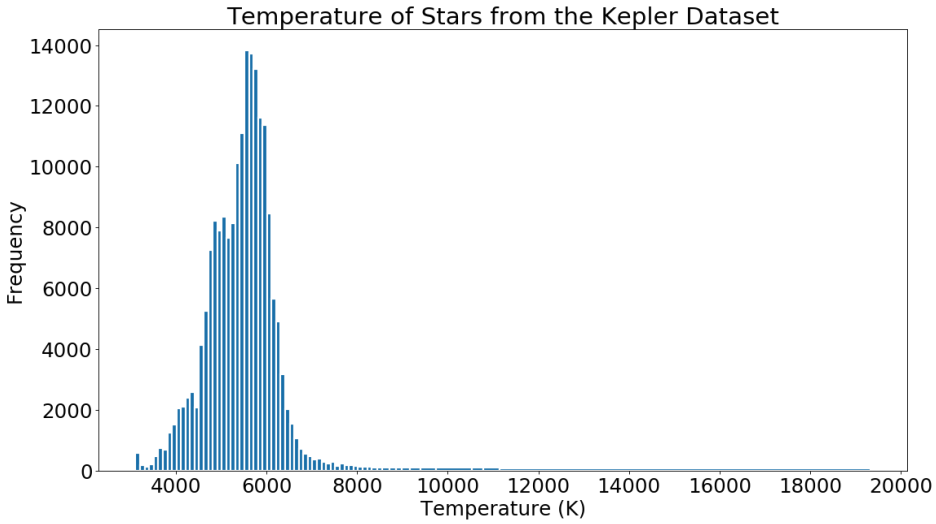
$$r_i(T) = 6.2817 \times 10^{-1} T^3 - 1.23515 \times 10^3 T^2,$$
$$r_o(T) = 1.52 T^3 - 2.98875 \times 10^3 T^2,$$
$$r_*(T) = 1.8395 \times 10^5 T + 3.6169 \times 10^8, \quad (2)$$

The equations determining the habitable zone were derived by calculating how far away from a star with temperature T, would a planet need to be to have a surface temperature between 225 and 350 degrees Kelvin.

2.3 *Kepler Dataset*

In order to form our dataset, we first pieced together the effective temperatures of each of the 191,451 stars observed in the Kepler 1 mission from the NASA Exoplanet archive [8]. The stellar effective temperature is essentially the temperature a black box would need to be to emit the same amount of electromagnetic radiation as the star in question.

2.4 *Bin Distributions*



Each of the 65 bins in our histogram were created with two specific conditions: each bin would have to contain at least 100 stars (though there were not enough stars remaining for the last bin), and no bin could have a range narrower than 100 degrees Kelvin. Defining this distribution was used in sections 3 and 4 to accurately model stellar parameters based off of the stars Kepler observed.

3. GEOMETRIC EFFICIENCY CALCULATION

3.1 *Overview*

Along with Kepler's detection efficiency, we also needed to account for the geometric factors that can prevent transit detection when using lightcurves. Because the Kepler Space Telescope can only see planets within its line of sight, planets orbiting in paths that don't cross the telescope's line of sight aren't seen. Therefore, many possible candidates for habitable exoplanets are lost.

3.2 *Equation*

To account for this, we calculated the geometric efficiency of the telescope. According to NASA [4], the geometric probability of observing a transit using a telescope is simply ($\frac{r_*}{\alpha}$) where $\propto$ is the orbital radius of a planet around the star. Because we are focusing on habitable planets, we only needed to calculate the geometric efficiency for planets within a star's habitable zone. To get this equation, we started with the equation ($\frac{r_*}{\alpha}$) and integrated with respect to α from r_i to r_o. This gave us the sum of all possible detection probabilities for a planet in the habitable zone. To get an average detection efficiency for a habitable planet around a star, we divided the equation by r_i-r_o. We then substituted the variables, r_*, r_o, and r_i to yield an equation for the probability of geometric detection as a function of temperature.

$$P_d(T) = \frac{r_*(T)\ln(\frac{r_o(T)}{r_i(T)})}{r_o(T) - r_i(T)} \tag{3}$$

3.3 *Calculation*

After averaging the geometric probabilities from every star in Kepler's dataset, we produced an approximation for the geometric efficiency of the Kepler Space Telescope.

$$\bar{P_d} \approx 6.4985 \times 10^{-3} \tag{4}$$

4. TRANSIT DETECTION EFFICIENCY CALCULATION

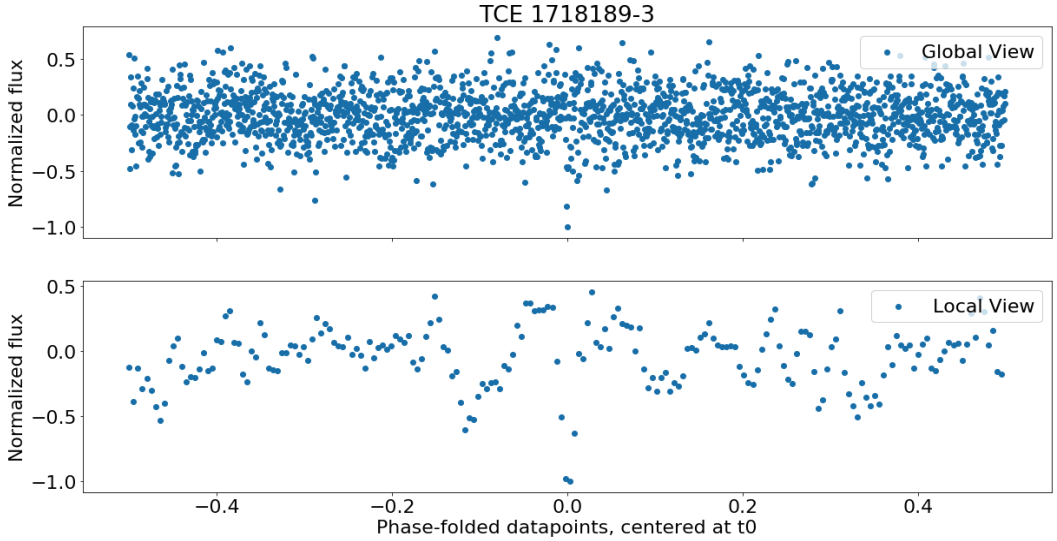
4.1 *Overview*

The Kepler Detection Efficiency was evaluated using a Monte Carlo approach, using simulated transits generated based on randomly chosen stellar effective temperatures proportional to the distribution found in section 2.4. When this simulated transit is fed into the trained machine learning model, the model's prediction for classifying the event theoretically yields the viability of the transit photometry method for that particular simulated system. If the model detects the transit, we can assume that the chosen stellar parameters are within the range of Kepler's detectability. If the model does not detect the transit, then we can assume that the chosen stellar parameters are outside the range of Kepler's detectability. The fraction of parameter vectors that could be detected over the total number of parameter vectors tested is considered the Kepler detection efficiency.

4.2 *Data Cleansing and Normalization*

In order to train the neural network for identifying fake transits, we cleansed and normalized labeled TCEs from the NASA Exoplanet Archive's Autovetter Planet Candidate Catalog for Q1-Q17 DR24 [8] using a process identical to the process stated in *Identifying Exoplanets with Deep Learning* [9]. In summary, low-frequency variability was removed from labeled TCEs by dividing each point outside of the main transit by a best-fit spline. Transits

outside of the main event were then removed along with >3σ outliers. Next, the data was phase-folded and then binned in two different modes: global and local views. The global view was generated by partitioning the data and obtaining the average value of each data point in these bins. The local view utilized overlapping bins specifically over the duration of the transit, resulting in a lightcurve that was more attentive to transit-specific trends. Once these views were generated, the data was ready to be used for training the algorithm.



4.3 *Machine Learning Algorithm*

A CNN was trained on the normalized events using the best model configuration from the paper *Identifying Exoplanets with Deep Learning* [9]. The model uses both the global and local input views from normalization and randomly chooses whether or not to flip each example along the horizontal axis during training. The model was trained over 50 epochs, using a batch size of 64, $\alpha = 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. To follow Shallue & Vanderburg's work as precisely as possible, training examples were still partitioned into three subsets: 80% training, 10% test, and 10% validation. Data was assigned the label 1 for confirmed planet candidates and the label 0 for both non-transiting phenomena and astrophysical false positives.

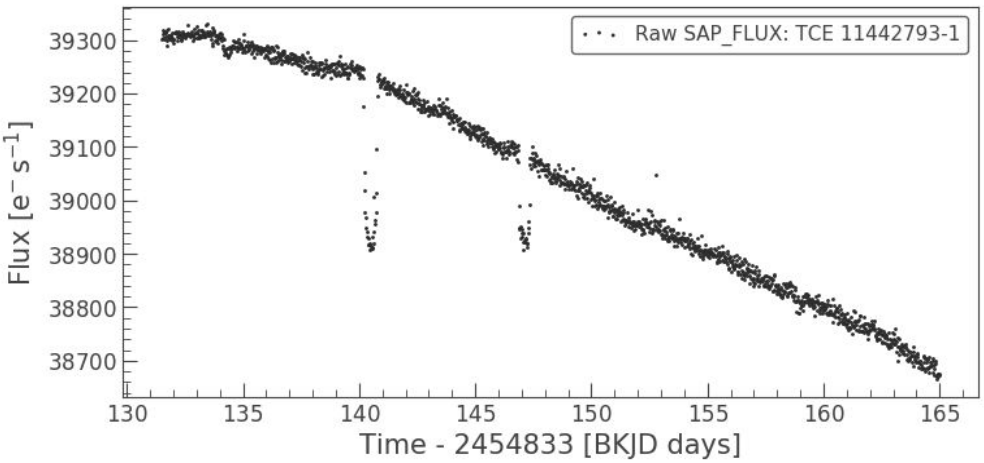
Despite following Shallue & Vanderburg's process as closely as possible, evaluation of our model yielded only 92.33% accuracy when classifying events in the test group. This most likely suggests an accidentally imperfect replication of their study. However, we believed the accuracy we achieved was still high enough to yield meaningful results in the remainder of our procedure.

4.4 *Simulated Transit Generation Algorithm*

In order to determine the limits of the machine learning algorithm in terms of astronomical parameters, simulated transits were generated using the library *batman*, described in the paper *BAsic Transit Model cAlculatioN in Python* [10] (Kreidberg, 2015). To simulate a transit assuming a circular orbit with 90º inclination using a uniform limb darkening model, the *batman* algorithm only requires the following parameters: orbital period, planet radius and orbital radius. We derived each of these from a given stellar temperature using the relationships

found in section 2.2. Values for planet radius and orbital radius were selected with uniform randomness within each derived range.

To introduce realistic noise to the simulated transits, we removed outliers greater than three MAD from the median from the lightcurve data for Kepler-90 (KIC 11442793). Outliers were replaced with gaussian noise using the mean and standard deviation of the lightcurve without outliers. We chose Kepler-90 because the algorithm used to remove its systematic trends visibly appeared to work well. Additionally, since Shallue & Vanderburg's algorithm assisted in the discovery of Kepler-90, we knew that the noise in that lightcurve did not interfere with proper detection of the system.



To complete a simulated transit, we subtracted the simulated flux from Kepler-90's noise, and fed this into the normalization process described in section 4.2. To generate the local view, a transit duration of 0.66 days was chosen, as equations using astronomical parameters yielded inconsistent and inaccurate results. This must be a fault of using improper relationships between our parameters and the transit duration, as hardcoding a set number of days is an uncomfortably poor programming practice. Derivation of relationships between astronomical parameters is one of the weakest portions of our procedure, which prompts further development.

4.5 *Kepler Detection Efficiency Calculation*

Through our experiment, the model detected 26,165 transits out of 998,400 total simulated transits generated off of stellar relationships derived in section 2.2 from effective temperatures matching the distribution described in section 2.4, yielding a Kepler detection efficiency of 2.62%. We assumed that our simulation converged at this result since a Kepler detection efficiency of 2.8% was evaluated from merely 9,000 samples, however; a more robust approach to approximating the error of this result must be taken in the future in order to verify this number.

5. CONCLUSIONS & DISCUSSION

5.1 *Results*

With the detection efficiency found, all that is left are simple calculations regarding the ratio of stars and planets within our galaxy to come to our conclusion. Plugging these numbers into the baseline equation in 2.1, our findings suggest that there are approximately

2.7113145e+12 planets within our galaxy that fit within our criteria for habitability. When averaged against the roughly 250 billion stars in the Milky Way, our number roughly simplifies to 11 habitable planets per star.

5.2 *Comparisons and Limitations*

Contrary to the estimates concluded by NASA and other studies, our findings tend to be considerably larger and thus leads us to revisit our attempts to pinpoint the concluding number. We have also taken into account the timing of our conclusions as a limitation, due to us attempting to reach this number before we could finish our research in full completion. However, another explanation of our conclusions is the geometric efficiency used in our equation, which is another aspect we aim to continue adjusting. Our plans for the future are to continue to critique this factor until we find that the accuracy of this efficiency to be satisfactory in regards to our conclusion.

5.3 *Looking Ahead*

Alongside learning of the limitations of efficiency of the Kepler Telescope, our team also acknowledges the advancement of technology over the years and have plans of reusing our research on the TESS Telescope, a newer telescope also created by NASA with similar strategies for finding habitable planets. This telescope focusing on a different part of the sky allows us new data that when used against our algorithm will allow an even more accurate number, due to the change of setting and increase of the efficiency detection factor provided. The exact number still remains unknown, but our aim to create an accurate and mathematically sound estimate still stands for the future. Furthermore, we did not consider the error bounds for any part of our calculations. This is an important part of any scientific endeavor, and for the future, we will definitely keep this in mind.

5.4 *Acknowledgments*

First and foremost, we want to thank Dr. Hawker for his immense help throughout this endeavor. Next, we thank the SIR department for guiding us throughout the process. Additionally, we wish to thank Terry Jones and Anthony Stuckey for their patience with us as we tried to utilize IMSA's mainframe for our research. We also want to acknowledge Andrew Vandenburg, one of the researchers behind the Google paper we often cited, for his willingness to guide us through the process of using and understanding his code. We also want to thank Anne Dattilo and Jacob Levine for additional assistance with using Chris Shallue & Andrew Vandenburg's code. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. This paper includes data collected by the Kepler mission. Funding for the Kepler mission is provided by the NASA Science Mission directorate. We thank Caltech and NASA for creating and managing the exoplanet archive and making it easily accessible to the public. Finally, we want to acknowledge Ed, Holly, and Aaron Fowler for providing and managing the computational resources used to yield our final results.

REFERENCES

[1] Borucki, W. J. (2010). Kepler space mission: Detection of Earth-size planets in the habitable zone of solar-like stars. 2010 IEEE Aerospace Conference. doi: 10.1109/aero.2010.5447039
[2] NASA, Europe Explore Joint Mission to Outer Planets. (2008). Physics Today. doi: 10.1063/pt.5.022133
[3] Boruki.
[4] 2017, About Transits, *NASA Kepler and K2 Learn More Page*, https://www.nasa.gov/kepler/overview/abouttransits (August 3, 2017)
[5] Schwamb, et al., (2012). Planet Hunters: Assessing the Kepler Inventory of Short Period Planets. *The Astrophysical Journal,* 3. doi: https://doi.org/10.1088/0004-637X/754/2/129
[6] Seager, S., & Mallen-Ornelas, G. (2003). A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. *The Astrophysical Journal*, 585:1038-1055. doi: https://doi.org/ 10.1086/346105
[7] Seager, S., & Mallen-Ornelas, G.
[8] Catanzarite, J. H. (2015). Autovetter Planet Catalog for Q1-Q17 Data Release 24 (KSCI-19090-001), Tech. rep. https://exoplanetarchive.ipac.caltech.edu/docs/KSCI-19091-001.pdf
[9] Shallue, C. J., & Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning. *The Astronomical Journal*, 155 (2), 94. doi: https://doi.org/10.3847/1538-3881/aa9e09
[10] Kreidberg, L. (2015). batman: BAsic Transit Model cAlculatioN in Python. *Publications of the Astronomical Society of the Pacific,* 127 (957), 1161–1165. doi: https://doi.org/10.1086/683602