Proceedings of the 2023 International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/10/20230343

Statistical analysis of terrestrial planets formation and observational bias

Steven Shangtong Ke

Sha Tin College, Sha Tin, NT, Hong Kong, 999077, China

kes2@stconline.edu.hk

Abstract. When investigating and researching into extrasolar planets, astronomers must deal with the inherent selection bias that arises because of limitations in technology or resources. Ideally, astronomers should aim to have a database with unbiased samples from across the universe. The ability for astronomers to efficiently use equipment will rely on statistical models employed to discover regions in which astronomical observations for extrasolar planets are lacking. Previous researchers have focused on the observation of Hot-Jupiters, Super Earths etc. This paper will be specific in exploring the selection bias involved with the detection of extrasolar terrestrials. In addition, the paper will investigate the probability of finding such terrestrial planets depending on the characteristics of exoplanet host stars. Ideally, the distribution of terrestrial planets should be random if the Earth is not considered special. The paper will also evaluate this hypothesis and determine whether there is any relationship between detection probability for exoplanets and their distance from Earth. Statistical methods such as logit regression and kernel estimations will be used in Stata to analyse data from the exoplanet encyclopedia. Hopefully, this paper will be able to provide more insight into the distribution of terrestrial planets and bias involved in detecting such planets.

Keywords: statistical analysis, terrestrial planets, solar properties, observational bias.

1. Introduction

In recent years, there has been a heightened interest in the study for extrasolar planet. The detection of these extrasolar planets is motivated by one of the key areas of study for many scientists. The field of detecting extrasolar planets, specifically new earths captivate both the media and astronomers [1]. There are many methods with which these types of planets can be detected, but finding a new earth starts with the detection of extrasolar planets. Methods such as Transits and radial velocity method are used commonly [2, 3]. However, with each method comes limitations involving the types of planets they are good at detecting. Large number of resources are being put into cataloguing the nearest and brightest stars, especially with TESS [3]. The exoplanet encyclopedia provides data on 5356 confirmed exoplanets in 3943 planetary systems [4]. Even with a vast amount of data, there will be areas in the database where selection bias has reduced the accuracy of a consensus. As astronomers only have access to light and more recently gravitational waves which reach Earth, there are a lot of technological limitations in the detection of exoplanets. The research objective of this paper is to use statistical methods to estimate the selection bias in observing terrestrial exoplanets and hopefully inform Astronomers about key areas of observation which are likely to be more useful. Previously there has been a focus on frequency and

DOI: 10.54254/2753-8818/10/20230343

probability of planetary populations, including the investigation of types of planets expected in average stellar systems [5, 6]. However, as seen from the summary of statistical analysis in exoplanets, there hasn't been a lot of modelling for the quality of the data used within existing research [5, 6]. Software such as EXONEST have used Bayesian methods to explore existing exoplanets, but STATA is also commonly used for computing statistics as well [7]. This paper would like to use linear regression, logit regression and maximum likelihood estimations to draw some conclusions regarding the classification of terrestrial exoplanets as a bonus as well [8]. In this respect, other researchers have also used similar statistical methods to find patterns in gaseous exoplanet formation [5]. A good example of using data as evidence for modelling astrophysical relationships would be the research completed using CARMENES [9]. Relationships between host star metallicity, effective temperature and stellar mass were confirmed using data modelling during the research [9]. Finally, the paper will begin by defining terrestrial planets, including the use of previous research regarding planet densities [10].

2. Methods

To classify extrasolar terrestrial data from the encyclopedia, the paper will make use of density parameters of 3.6 to 13.4 grams per centimeter cubed (g/cc) already determined by other researchers [4, 10]. Density can be calculated by given data on planetary mass and planetary radius measured in Jupiter masses and Jupiter Radius. A simple proof of method is as follows:

$$\rho = \frac{m * M_{jup}}{\frac{4}{3} * \pi * \left(r * R_{jup}\right)^3} = \frac{m * M_{jup}}{r^3 * \frac{4}{3} * \pi * R_{jup}^3} = \frac{m * M_{jup}}{r^3 * V_{jup}} = \frac{m}{r^3} * \rho_{jup} \quad (1)$$

Therefore, dividing the mass from the encyclopedia by the radius cubed will provide the relative density of the exoplanet compared to Jupiter, allowing for the density of the exoplanet in terms of g/cc to be calculated.

Once data for density of exoplanets have been processed, a dummy is created to find all the extrasolar planets with densities within the range of 3.6 g/cc and 13.4 g/cc. 317 planets were found to be within the rocky planet range, significantly more than the 19% of 570 planets found by previous researchers [10]. One of the reasons for this significant disparity is due to the increase in sample size from 510 to 1322 exoplanets with both radius and mass measured on the database.

Table 1. summary statistics.

Variable	Observation	Mean	Standard Deviation	Minimum	Maximum
rocky_dummy	5,346	0.0592967	0.2362012	0	1
star_distance	5,029	696.2306	1172.15	1.295	11000
star_metallicity	4,304	-.0050088	0.2011382	-1.08	0.61
star_mass	4,774	0.9523254	0.4825216	0 .011	20
star_radius	4,449	1.542683	3.701767	0.0083	88.5
star_age	2,882	4.206375	2.541629	0.00002	15
star_teff	4,653	5434.41	1547.392	48.91	42000
logstardistance	5,029	5.611297	1547.392	0.2585107	9.305651
logstarteff	4,653	8.574648	0.2383473	3.889982	10.64542

To clarify the source of the data that the paper will be using, Table 1 provides a summary of all the variables taken from the EU catalogue for exoplanets [4]. The dummy mentioned earlier is also summarized within the table. The table also shows a summary regarding the log of star distance, which will be useful for statistical analysis later.

2.1. Solar properties regression

With the data and classification of terrestrial planets completed, a regression of density against different solar properties will yield any correlations within the data. Suppose a linear regression for equation:

DOI: 10.54254/2753-8818/10/20230343

$$Y = \boldsymbol{\alpha X} + \varepsilon \tag{2}$$

The paper will investigate different solar properties illustrated in Table 1 as **X**, and probability for terrestrial planet appearance as Y. Epsilon represents the value of random noise. $\boldsymbol{\alpha}$ is the coefficient of regression, which will be used to express any relationship between **X** and Y. The code for the regression is written in STATA. To ensure accuracy, the paper will be using both logit and probit regression.

Because the values used for regression is a binary dummy, logit regression is the most suitable method [8]. The following will be a brief introduction to the logit regression used in the paper. **X** is a vector of all the different solar properties the paper will be regressing, and $\boldsymbol{\alpha}$ is a vector of all the coefficients for each corresponding variable. Here is the probability density function used:

$$logit(\boldsymbol{X}) = \frac{e^{\boldsymbol{\alpha X}}}{1 + e^{\boldsymbol{\alpha X}}} \tag{3}$$

Using equation (3) it is possible to regress the probability of formation of terrestrial planets.

2.2. Modelling bias

The paper will also investigate the selection bias involved with exoplanets in general. By modeling the frequency of stars found with exoplanets and compare the host star's distance to Earth. The graph created will have a corresponding function $\hat{p}(d)$, in which frequency of star found away from Earth will vary with d (host star's distance to earth).

If the probability of terrestrial planets is modelled previously to be $\hat{Y}$, then the function $\hat{p}(d)$ should have no correlation with $\hat{Y}$ without any observational bias. Imagine a sphere of radius d with the Earth at the center, $\hat{Y}$ can also be expressed as the total frequency of stars with exoplanets detected within the sphere divided by the number of host stars with terrestrial exoplanets. Theoretically, if the distribution of extrasolar terrestrial planets is random, then the frequency of stars detected and d should produce a cubic relationship. Hypothetically the function $\hat{p}(d)$ should vary with d^2 as $\hat{p}(d)$ is the frequency of exoplanet host stars found on the surface of the sphere d. To plot the graph, a frequency density histogram will be used along with a kernel density estimation.

The graph should also provide information about the function of detection probability $\hat{g}(d)$. If we normalize the probability function $\hat{p}(d)$ such that $\hat{g}(d)$ is within a unit sphere, then we obtain:

$$\hat{g}(d) \propto \frac{\hat{p}(d)}{d^2} \tag{4}$$

To test whether there is a selection bias for terrestrial planets regarding distance from earth, a regression can be completed on $\hat{Y}$ and $\hat{g}(d) * d^2$. If $\hat{Y}$ and $\hat{g}(d) * d^2$ has no correlation at all, then the detection probability $\hat{g}(d)$ and $\hat{Y}$ do not change with d. Because the probability of terrestrial planets should not change with distance from earth, no correlation will suggest that the dataset has no biases. On the other hand, if a relationship does appear, selection bias can be observed and modelled.

3. Results

3.1. Kernel plot of frequency of being rocky

After processing the data, all the graphing and plotting could begin in STATA with a little coding.

DOI: 10.54254/2753-8818/10/20230343

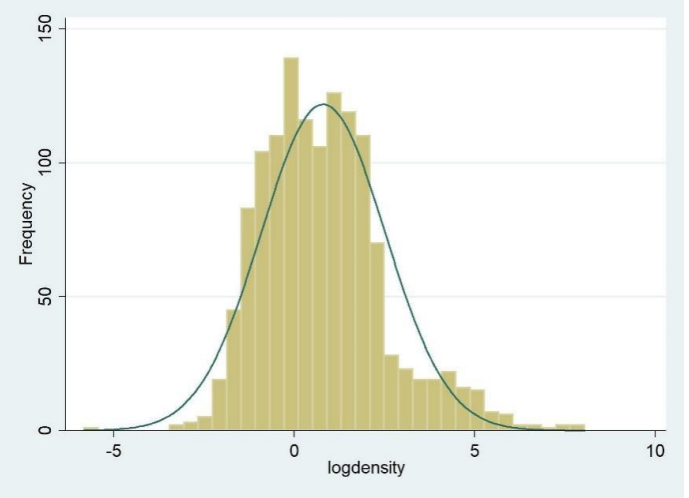

Figure 1. histogram of frequency of exoplanets against their log density.

Figure 1 shows a log scale of density against frequency on a 35-bin histogram plot, with a binomial distribution curve overlayed. There is a clear dip in the data at the 0.56 mark, which is around log(3.6). On the other end, the slope begins to steep after log(13.4), suggesting the upper bound is also within range. Overall, the result of this graph reflects a similar trend to previous research and the bounds determined are still viable [10].

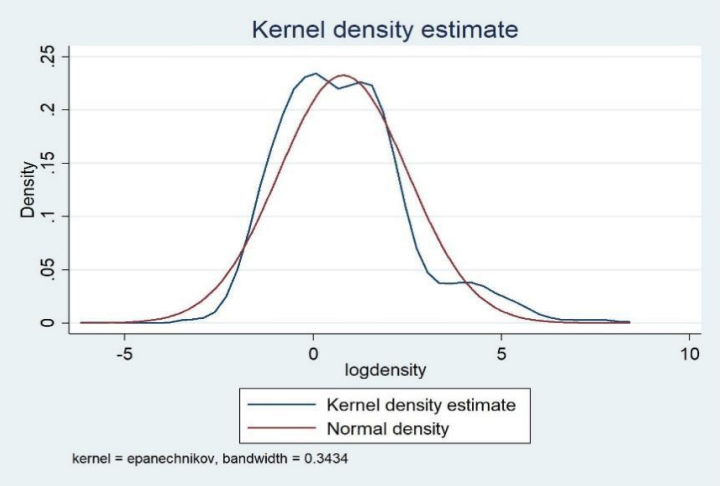

Figure 2. kernel density estimate of density against log density of exoplanet.

Figure 2 shows the Kernel density estimate for log density against frequency density. There is an even clearer trend that corresponds with the upper and lower bound of earth-like planets determined between 3.6 g/cc and 13.4 g/cc. The first peak of the graph reflects the average density of gaseous planets and the final peak at the log density of 5 may refer to any outlying brown dwarf-like planets [10]. One difference in the new distribution against previous research is that the frequency density of terrestrial planets is significantly higher compared to gaseous planets. The change may highlight a selection bias, as astronomers shift more resources into using methods to accurately find smaller terrestrial planets.

3.2. Correlation matrix and its facts

There may already be relationships between certain solar properties. A correlation matrix will be used to illustrate these relationships. These relationships could have an impact on the probability of terrestrial planets forming.

DOI: 10.54254/2753-8818/10/20230343

Table 2. correlation matrix between solar properties. Standard errors are in parenthesis.

	star_distance	star_metallicity	star_mass	star_radius	star_age	star_teff
star_distance	1					
star_metallicity	-0.0874 (0.0000)	1				
star_mass	-0.1184 (0.0000)	0.2058 (0.0000)	1			
star_radius	-0.0603 (0.0001)	-0.0894 (0.0000)	0.4214 (0.0000)	1		
star_age	-0.0637 (0.0007)	-0.0467 (0.0149)	-0.1877 (0.0000)	-0.0545 (0.0049)	1	
star_teff	0.2952 (0.0000)	0.108 (0.0000)	0.4251 (0.0000)	-0.0623 (0.0000)	-0.1298 (0.0000)	1

As seen in Table 2, a correlation matrix created in STATA with significance values under each comparison, many of the solar properties have correlations. For example, assuming that most observed systems are main sequence stars, there is a relationship between star radius and star mass. With a 0.4214 correlation, as star mass increases, the radius and therefore volume will tend to increase as well. In other words, the mass and radius trend describe the same physical relationship. Additionally, a study using data from CARMENES also confirms a correlation between effective star temperature, star metallicity and stellar mass [10]. Thus, the coefficients in Table 2 further provides some evidence towards the interconnectedness of such star properties.

With an $\alpha = -0.281$ from logit regression and $\alpha = -0.137$ from probit regression, there is a negative correlation between distance from the host star and the likelihood of a terrestrial planet. From previous research, it is known that smaller and denser planets are found closer to the sun with lower orbital periods [5]. Terrestrial planets are normally a lot denser than gaseous planets, as defined previously in the paper and in other literature [10]. Therefore, a negative correlation between the probability of a terrestrial planet appearing and the distance from the star is expected. In addition, the relationship from Table 3 between stellar mass, stellar radius, star effective temperature and star metallicity also play a similar role in affecting the probability of terrestrial planet appearance.

3.3. Logit regression for the liklihood of being rocky

Table 3. Probit and logit Regression for coefficient α from equation (2) for multiple solar properties, with confidence rates. Standard errors are in parenthesis, *** is p<0.01, ** is p<0.05 and * is p<0.1.

	Logit model	Probit model
VARIABLES	rocky_dummy	rocky_dummy
logdisttostar	-0.281*** (0.075)	-0.137*** (0.035)
star_metallicity	-1.971*** (0.479)	-1.073*** (0.260)
star_mass	4.172*** (1.062)	2.082*** (0.571)
star_radius	-1.365*** (0.399)	-0.631*** (0.174)
star_age	-0.029 (0.033)	-0.019 (0.018)

DOI: 10.54254/2753-8818/10/20230343

Table 3 (continue)

logstarteff	-5.138*** (1.242)	-2.786*** (0.692)
Constant	39.016*** (9.885)	21.194*** (5.513)
Observations	1,248	1,248
log likelihood	-378.9	-378.9
chi-squared	89	89

In Table 3, solar properties including distance of exoplanet to host star, star metallicity and star's effective temperature provide a good correlation for investigation. The standard errors for star distance, star metallicity, effective star temperature and star radius are rather minimal with 99% confidence for both logit and probit regressions.

As an example, a calculation of how probability changes for stellar mass with an $\alpha = 4.172$ assuming ceteris paribus. Taking the mean for all significant values in Table 2 will provide a vector for $\boldsymbol{X_{\mu}}$. If $\boldsymbol{\alpha_{\mu}}$ is a vector of the coefficients calculated, plugging into equation (3) will provide the probability of terrestrial planets appearing for mean stellar mass m_{μ}. After the dot product, including the constant variable, the probability calculated from equation (3) is around $8.58 * 10^{-3}$. If a standard deviation is summed onto the mean of the stellar mass, we will obtain $6.087 * 10^{-2}$ for the probability. Thus, the probability increases by 6 times when the mass changes by one standard deviation from the mean, ceteris paribus. The result makes sense as the standard deviation of stellar mass is almost 50% the mean of stellar mass.

3.4. Observational biasness

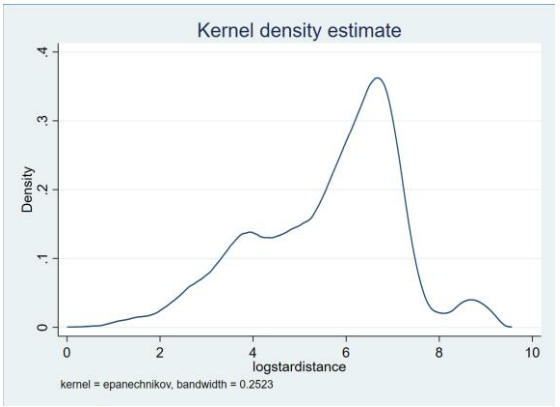

Figure 3. kernel density estimate of frequency density against log distributed star distance from earth (parsec/pc).

In Figure 3, the graph and function for $\hat{p}(d)$ is shown through kernel estimation. Up until e^7or 1097 pc away from Earth, the relationship is somewhat like a quadratic curve sloping upwards. However, after 1097 pc there is a clear dip, and the trend varies away from a quadratic curve. The estimation provides information regarding star distribution, which suggests that the current data from the encyclopedia does not have enough samples for farther star systems with exoplanets. If there was no bias, the graph would look more like a quadratic. In addition, at e^8 or around 3000 pc away from earth, there a clear lack of data for exoplanets. The main reason behind these disparities should include technological limitations in telescopes or surveys including Kepler's in 2010 [1]. The farther away star systems are, the less likely there are enough accurate telescoping resources to observe these planets. Methods of detection for these exoplanets, including transits, are also less likely to be observable the farther exoplanets are from Earth. Thus, it is reasonable for figure 3 to have such a correlation.

DOI: 10.54254/2753-8818/10/20230343

Table 4. Ordinary least square regression for $\hat{Y}$ and $\hat{g}(d) * d^2$. Standard errors are in parenthesis, *** is p<0.01, ** is p<0.05 and * is p<0.1.

	OLS
Variables	yhat
pdensity	0.094***
	(0.029)
Constant	0.099***
	(0.003)
Observations	1,105
R-squared	0.009

Table 4 shows the coefficient between pdensity, which is $\hat{g}(d) * d^2$ and yhat ($\hat{Y}$) to be 0.094 with a confidence of 99%. The positive relationship between the probability of terrestrial planets and the detection probability of exoplanets suggests that the farther away the star system is from earth, the "more likely" terrestrial planets are. If there is no bias at all, $\hat{Y}$ should not change with d and detection probability of exoplanets should not vary with $\hat{Y}$ at all.

3.5. Discussion

The main point of discussion for the results of the paper is the distribution of stars across the galaxy. For the relationship between the distance of the host star to Earth and the frequency of these stars d, a logscale was used in the paper. This could hint at the possibility that the distribution of stars and thus exoplanets is non-Euclidean. The distribution of stars and exoplanets will affect the fundamental assumptions within the paper, and the ability for observational tools to fairly detect exoplanets in the universe. The implications of such a distribution would be far-reaching, but currently it requires more research and verification.

Selection bias may also contribute to a general limitation of the evidence. Due to the use of the transit method and radial velocity method to find the radius and the mass of exoplanets respectively, there exists a selection bias between the types of planets these methods are good at finding. Advances in observational technology should allow for a better distribution of exoplanets across all distances in databases. Currently, more care should be taken when processing data from exoplanet databases.

4. Conclusion

This paper has provided updated evidence regarding factors affecting the formation of terrestrial planets. Furthermore, the paper modelled and tested whether data from the encyclopedia for exoplanets is biased. The paper should also contribute as a starting point for future researchers with more available data to statistically analyse earth-like planets. Information regarding the distribution of discovered stars with exoplanets should shed light on where astronomers should be directing their resources and attention to observe exoplanets in the galaxy. If astronomers can amend the data at regions between e^4 and e^8pc away from earth, the exoplanet database will become less biased for future use.

In the short term, more astronomical observations about extrasolar planets will aid scientists in discovering new earths. Similar statistical models for databases and exoplanet types will increase the efficiency of exoplanet detection. Hopefully, the paper will inspire more researchers to create more accurate models regarding database incompleteness to help aid observers and researchers. In the long term, as technology for spacefaring advances, it is important for humans to research other habitable systems and understand the formation of terrestrial exoplanets. Through understanding these systems people can investigate the possibility of leaving the solar system, providing clues to the formation of our own solar system, yielding interesting conclusions regarding the emergence of life on Earth.

DOI: 10.54254/2753-8818/10/20230343

References

[1] Borucki W J et al. 2010 Kepler planet-detection mission: introduction and first results *Science* 327 p 977-80.
[2] Ricker G R et al. 2014 The transiting exoplanet survey satellite *Journal of Astronomical Telescopes Instruments and Systems* 1 p 14003.
[3] Reiners A, Bean J L, Huber K F, Dreizler S, Seifahrt A and Czesla S 2010 Detecting planets around very low mass stars with the radial velocity method *The Astrophysical Journal* 710 p 432.
[4] The extrasolar planets encyclopedia http://exoplanet.eu/catalog/ update: April 6, 2023 (5357 planets).
[5] Zhu W and Dong S 2021 Exoplanet statistics and theoretical implications *Annu. Rev. Astron. Astrophys* 59 p 291-336.
[6] Emsenhuber A, Mordasini C, Burn R, Alibert Y, Benz W and Asphaug E 2022 The new generation planetary population synthesis (NGPPS) II *A&A* vol 656 (1) A70.
[7] Knuth K H, Placek B, Angerhausen D, Carter J L, D'Angelo B, Gai A D and Carado B. 2017 EXONEST: The Bayesian Exoplanetary Explorer *Entropy* vol 19 (10) p 559.
[8] Greene H W 2002 *Econometric Analysis* Pearson Education New York City p 896
[9] Passegger V M et al. 2018 The CARMENES search for exoplanets around M dwarfs – photospheric parameters of target stars from high-resolution spectroscopy *A&A* vol 615 A6.
[10] Odrzywolek A and Rafelski J 2018 Classification of Exoplanets According to Density *Acta Physica Polonica B* vol 49 p 1917-22.