Pros and Cons of Classification of Exoplanets: in Search for the Right Habitability Metric

Margarita Safonova, Snehanshu Saha, Jayant Murthy, Suryoday Basak, Madhu Kashyap, Surbhi Agrawal, Kakoli Bora

Since time immemorial, humanity was looking at cosmos and believing other worlds being out there, inhabited with other or, may be same, beings like us. Indian ancient texts talk about travelling to other worlds in `bodily form’ (inscription on the iron pillar at Qutub Minar, probably left in 4th century BC). Ancient Greeks also believed in the existence of other planetary bodies with beings living on them (with mentions dating as far back as 6th century BC: Thales of Miletus and Pythagoras). With our technological advances, we are continuing this same quest – the quest for the habitable planet, ultimately, the second Earth; or at the least, for the answer to the question of whether we are alone in the Universe. Currently, we already know about the existence of thousands of exoplanets, and the estimates of the actual number of planets exceed the number of stars in our Galaxy alone by orders of magnitude (both bound and free-floating); small rocky planets, super-Earths, the most abundant type. Our interest in exoplanets lies in the fact that, anthropically, we believe that life can only originate and exist on planets, therefore, the most fundamental interest is in finding a habitable planet - the planet with life on it. This quest can be broadly classified into the following: looking for Earth-like conditions or the planets similar to the Earth (Earth similarity), and looking for the possibility of life in a form known or unknown to us (habitability). But what is habitability? Is it the ability of a planet to beget life – a potential habitability? Or is it our ability to detect it: a planet may host life as we know it, in other words, be inhabited, but we will not detect it unless it evolved sufficiently to change the environment on a planetary scale. In both cases, the only comparison for recognition is our planet, therefore, we are looking for the terrestrial likeness in exoplanets.

With a constantly increasing number of discovered exoplanets and the possibility that stars with planets are a rule rather than an exception, it became possible to begin characterizing exoplanets in terms of planetary parameters, types, populations and, ultimately, in the habitability potential. This is also important in understanding the formation pathways of exoplanets. But, since complete appraisal of the potential habitability needs the knowledge of multiple planetary parameters which, in turn, requires hours of expensive telescope time, it became necessary to prioritise the planets to look at, to develop some sort of a quick screening tool for evaluating habitability perspectives from observed properties. Here, the quick selection is needed for a long painstaking spectroscopic follow-up to look for the tell-tales of life – the biosignatures – atmospheric gases that only living organisms produce in abundance. It can be oxygen, ozone, methane, carbon dioxide or, better, their combinations (see e.g. Safonova et al. 2016; Krissansen-Totton et al. 2018, and references therein). For all the upcoming space missions dedicated to the search of life in the Universe: PLATO, Euclid, JWST, etc., we need to make a list of our preferred candidates, so that this quest hopefully can be completed within our lifetimes. It is estimated that one in five solar-type stars and approximately half of all M-dwarf stars may host an Earth-like planet in the habitable zone (HZ). Extrapolation of Kepler's data shows that in our Galaxy alone there could be as many as 40 billion such planets (e.g. Borucki et al. 2010; Batalha et al. 2013; Petigura et al. 2013). And it is quite possible that soon we may actually detect most of them. But with the ultimate goal of a discovery of life, astronomers do not have millennia to quietly sit and sift through more information than even pentabytes of data. Obtaining the spectra of a small planet around a small star is difficult, and even a large-scale expensive space mission (such as e.g. JWST) may be able to observe only about a hundred stars over its lifetime (e.g., Turnbull et al. 2012).

For that purpose, several assessment scales have been introduced: a concept of the Habitable Zone (HZ) — a range of orbital distances from the host star that allows the preservation of the water in liquid state on the surface of a planet (Kasting et al., 1993); Earth Similarity Index – an ensemble of planetary physical parameters with Earth as reference frame for habitability, or Planetary Habitability Index (PHI), based on the biological requirements such as water or a substrate (Schulze-Makuch et al. 2011); habitability index for transiting exoplanets (HITE) based on the certain limit of planetary insolation at the surface (Barnes et al. 2015). Our group has developed an index applicable to small planets – Mars Similarity Index (MSI) -- as potential planets to host extremophile life forms (Kashyap et al. 2017). Habitability may also be viewed as probabilistic measure, in contrast to the binary definition of, say, being in the HZ or not, and such approach requires optimization classification methods that are part of machine learning (ML) techniques. Thus, we have introduced a Cobb-Douglas Habitability Score – an index based on Cobb-Douglas habitability production function (CD-HPF), which computes the habitability score by using measured and estimated planetary parameters (Bora et al. 2016), and recently extended it to include a statistical ML classification method (XGBoost) used for supervised learning problems, where the training data with multiple features are used to predict a target variable (Saha et al. 2018).

However, all classification strategies have caveats, and some (e.g. Tasker et al. 2017) reject the exercise entirely on the basis of impossibility to quantitatively compare habitability, and on the idea that pretending otherwise can risk damaging the field in the eyes of the public community. In addition, some researchers believe that the priority for the exoplanet and planetary science community is to explore the diversity of exoplanets, and not to concentrate on exclusions.

The first qualitative scale for habitability was the concept of a HZ – it assumes once a planet is in the HZ, it is potentially habitable. However, such criterion is binary, and we know that e.g. our Moon is inside the HZ, and is a rocky planetary body, but definitely not potentially habitable for our kind of life. Earth itself is located on the very edge of the HZ (making it marginally habitable) and will get out of it in the next 1-3 billion years. Mars is technically inside the HZ, and Venus once was. Titan, on the other hand, is totally outside the HZ but may host a life, albeit dissimilar to ours. Besides, recent discoveries of free-floating planets (planets without the host star where the concept of a HZ cannot apply) brought back the interest in their potential habitability that was first addressed in 1999 (Stevenson, 1999).

Coming to more quantitative assessments, HITE predicted that planets that receive between 60–90% of same amount of insolation as Earth, are likely to be habitable. It however assumes only circular orbits and the location inside HZ, which again refers back to mostly Earth similarity; besides our Solar System has a unique feature of very low ellipticities. TRAPPIST-1 and Proxima b having high ellipticities, and it was estimated that even high e orbits can have low effect on planetary climate provided they are in a certain resonance (p=0.1 preferably, where p is the ratio of ......).

The ESI is based on the well-known statistical Bray-Curtis scale of quantifying the difference between samples, frequently used by ecologists to quantify differences between samples based on count data. However, most multivariate community analyses are about understanding a complex dataset and not finding the ``truth", meant in a sense of ``significance". Thus, it may not be enough to understand a complex hierarchy of classification. But since all we know is the Earth-based habitability, our search for habitable exoplanets (an Earth-like life clearly favoured by the Earth-like conditions) has to be by necessity anthropocentric, and any such indexing has to be centred around finding Earth-like planets, at least initially. But Earth may not be the ideal place for life, and the concept of a super-habitability was introduced in 2014 (Heller and Armstrong, 2014). Though this concept got rid of a HZ limits admitting the tidal heating as a possible heat source, it still assumed the necessity of liquid water on the surface as a prerequisite for life, preferably as a shallow ocean with no large continuous land masses. Recent simulations, however, has shown that too much water is not good for the detectability of exo-life (Desch et al. 2018). Exoplanets without land would have life with much slower biogeochemical cycles and oxygen in the atmosphere would be indistinguishable from the one produced abiogenically. The question now shifts to the definition of habitability as our ability to detect it - if we cannot get to the planets which may have life not on the surface, they are as good as uninhabited.

The search for life on the planets outside the Solar System can be broadly classified into the following: looking for Earth-like conditions or the planets similar to the Earth (Earth similarity), and looking for the possibility of life in a form known or unknown to us (habitability). The two frequently used indices, ESI and PHI, describe heuristic methods to score similarity/habitability in the efforts to categorize different exoplanets or exomoons. ESI, in particular, considers Earth as the reference frame for habitability and is a quick screening tool to categorize and measure physical similarity of any planetary body with the Earth. The PHI assesses the probability that life in some form may exist on any given world, and is based on the essential requirements of known life: a stable and protected substrate, energy, appropriate chemistry and a liquid medium. Bora et.al (2017) proposed a different metric, a Cobb-Douglas Habitability Score (CDHS), based on Cobb-Douglas habitability production function (CD-HPF), convex optimization techniques and constrained behavior of the optimizing model drawing inspiration from the earlier work (Ginde, G et. al., 2016 and Saha, S et.al, 2016) which computes the habitability score by using measured and calculated planetary input parameters. The metric, with exponents accounting for metric elasticity, is endowed with verifiable analytical properties that ensure global optima, and is scalable to accommodate finitely many input parameters. The model is elastic, does not suffer from curvature violations and, as the authors discovered, the standard PHI is a special case of CDHS. Computed CDHS scores are fed to K-NN (K-Nearest Neighbour) classification algorithm with probabilistic herding that facilitates the assignment of exoplanets to appropriate classes via supervised feature learning methods, producing granular clusters of habitability. The proposed work describes a decision-theoretical model using the power of convex optimization and algorithmic machine learning. Saha et al. (2018) expanded previous work by Bora et al. (2017) on using Machine Learning algorithm to construct and test planetary habitability functions with exoplanet data. This time they analyzed the elasticity of their Cobb-Douglas Habitability Score (CDHS) and compared its performance with other machine learning algorithms. They demonstrate the robustness of their methods to identify potentially habitable planets from exoplanet dataset. Given our little knowledge on exoplanets and habitability, the results have limited value now. However, their methods provide one important step toward automatically identifying objects of interest from large datasets by future ground and space observatories. Therefore, our work provides a logical evolution from the previous work by Bora et.al (2017). CDHS contributes to the Earth similarity concepts where the scores have been used to classify exoplanets based on their degree of similarity to Earth. However Earth similarity is not equivalent to exoplanetary habitability. Therefore, Saha et. al. (2018) adopted another approach where machine classification

algorithms have been exploited to classify exoplanets into three classes: nonhabitable, mesoplanets and psychroplanets (originally adopted by the Planetary Habitability Laboratory (PHL), http://phl.upr.edu). While this classification was performed, CDHS was not used at all, rather discriminating features from the PHL-EC were used. This is fundamentally different from CDHS-based Earth similarity approach where explicit scores were computed. Therefore, it was pertinent and remarkable that the outcome of these two fundamentally distinct exercises reconcile. This reconciliation approach is the first of its kind and fortifies CDHS, more than anything else. Moreover, this convergence between the two approaches is not accidental.
Essentially, the ESI score gives non-dynamic weights to all the different planetary (with no trade-off between the weights) observables or calculated features considered, which in practice may not be the best approach, or at least, the only way of indicating habitability. It might be reasonable to say that for different exoplanets, the various planetary observables may weigh each other out to create a unique kind of favorable condition. For instance, in one planet, the mass may be optimal, but the temperature may be higher than the average of the Earth, but still within permissible limits (like Venus); in another planet, the temperature may be similar to that of the Earth, but the mass may be much lower. By discovering the best combination of the weights (or, as we call it,elasticities) to maximize the resultant score, to the different planetary observables, we are creating the metric which presents the best case scenario for the habitability of a planet.

The essence of the CD-HPF, and consequently, that of the CDHS is indeed orthogonal to the essence of the ESI or BCI. The argument is not in favor of the superiority of our metric, but for the new approach that have been developed. There should actually be various metrics arising from different schools of thought so that the habitability of an exoplanet may be collectively determined from all these. Such a kind of adaptive modeling has not been used in the context of planetary habitability prior to the CD-HPF.

The CD-HPF reconciles with the machine learning methods that have been used to automatically classify exoplanets. It is not easy to propose two fundamentally different approaches (one of which is CDHS) that lead to a similar conclusion about an exoplanet. While the CDHS provides a numerical indicator (in fact, the existence of one global optima shouldn't be a concern at all but rather a vantage point of the model that thus eliminates the possibility of computing scores arbitrarily), the machine classification bolsters the proposition by telling us automatically which class of habitability an exoplanet belongs to. Bora et al. explain that a CDHS close to 1 indicates a greater chance of habitability. The performance of machine classification is evaluated by class-wise accuracy. The accuracies achieved are remarkably high, and at the same time, it is observed that the values of the CDHS for the sample of potentially habitable exoplanets which are considered are also close to 1. Therefore, the computational approaches map Earth similarity to habitability. This is remarkable and non-trivial.

Despite recent criticism of the whole idea of exoplanetary ranking, we are sure that this field has to continue and evolve to use all available machinery of astroinformatics and machine learning. It might actually develop into a sort of same scale as stellar types in astronomy. It can be used as a quick tool of screening planets in important characteristics in search for potentially habitable planets for the follow-up investigations.

References:

Kasting, J. F., Whitmire, D. P., and Reynolds, R. T.,1993. Habitable zones around main sequence stars. *Icarus*, **101**, 108-108.

Schulze-Makuch, D., Méndez, A., Fairén, A. G., von Paris, P., Turse, C., Boyer, G., Davila, A. F., Resendes de Sousa António, M., Irwin, L. N., and Catling, D., 2011. A Two-Tiered Approach to Assess the Habitability of Exoplanets.

J. Krissansen-Totton el al., Disequilibrium biosignatures over Earth history and implications for detecting exoplanet life, *Science Advances* (2018), DOI: 10.1126/sciadv.aao5747

Saha et.al., 2018. Theoretical Validation of Potential Habitability via Analytical and Boosted Tree Methods: An Optimistic Study on Recently Discovered Exoplanets, arXiv:1712.01040 [astro-ph.EP]

Bora, K. et. al., 2016. CD-HPF: New Habitability Score Via Data Analytic Modeling. Astronomy and Computing, 17, 129-143

Saha, S. et. al., 2016. A novel revenue optimization model to address the operation and maintenance cost of a data center. Journal of Cloud Computing, 5:1, 1-23

Ginde, G. et.al, 2016. ScientoBASE: A Framework and Model for Computing Scholastic Indicators of Non-Local Influence of Journals via Native Data Acquisition Algorithms. J. Scientometrics, 107:1, 1-51

Safonova, M., Murthy, J., & Shchekinov, Y. A. 2016. Age aspects of habitability. International Journal of Astrobiology, 15, 93

Stevenson, D. J. 1999, Life-sustaining planets in interstellar space? Nature, 400, 32

Heller R. & Armstrong J. 2014. Superhabitable Worlds. Astrobiology, 14, 50-66.