Machine learning: investigando a relação entre atmosferas biogênicas, exoplanetas e sistemas planetários

Luander Bernardes[1,2,3] e Anna Carolina Martins[1,4]

[1]*Universidade de São Paulo*
[2]*Centro Universitário Estácio de Ribeirão Preto*
[3]*Instituto Federal de Educação, Ciência e Tecnologia de São Paulo*
[4]*Universidade Estadual de Campinas*

Resumo
A proposta da pesquisa foi estudar alguns sistemas planetários em busca de exoplanetas capazes de sustentar atmosferas biogênicas e entender o papel deles nesse contexto. Para essa tarefa, foi considerado o *Earth similarity index* (ESI), que é um índice de classificação de exoplanetas por similaridade com a Terra, porém, considerando uma perspectiva de modelagem multinível. A caracterização destes sistemas é importante, pois missões astronômicas, cujo objetivo seja identificar marcadores biológicos que denunciem a presença de vida, podem escolher alvos prioritários e eliminar alvos menos promissores. Foram estudados 72 sistemas extrassolares por meio de técnicas não supervisionadas e supervisionadas de *machine learning* com o objetivo de identificar a formação de *clusters* e investigar a relação multinível entre planetas e sistemas extrassolares. O trabalho demonstra que, provavelmente, uma ampla variedade de tipos de exoplanetas pode abrigar atmosferas aptas a serem estudadas remotamente, apesar de que esses resultados não consideram as constituições internas reais dos objetos estudados, já que elas são desconhecidas, o que impede uma reconstrução histórica do processo de evolução desses planetas. A abordagem multinível demonstra que aproximadamente 54% da variação do valor do ESI se deve ao efeito das condições do sistema planetário onde está localizado o exoplaneta em estudo.

Abstract
The research proposal was to study planetary systems with exoplanets capable of sustaining biogenic atmospheres and understand their role in this context. For this task, the Earth similarity index (ESI) coupled to a multilevel modeling perspective was considered, as it is an index for classifying exoplanets by similarity to Earth. The characterization of these systems is important because astronomical missions whose objective is to identify biological markers that reveal the presence of life will be able to choose priority targets and eliminate less promising targets. 72 extrasolar systems were studied using unsupervised and supervised Machine Learning techniques with the aim of identifying the formation of clusters and investigating the multilevel relationship between planets and planetary systems. The work demonstrates that a wide variety of types of exoplanets can probably harbor atmospheres capable of being studied remotely, although these results do not consider the real internal constitutions of the objects studied, as they are unknown, preventing a historical reconstruction of the evolution process of these planets. The multilevel approach demonstrates that approximately 54% of the variation in the ESI value is due to the effect of the conditions of the planetary system where the exoplanet under study is located.

Palavras-chave: exoplanetas, atmosferas biogênicas, aprendizado de máquina.
Keywords: exoplanets, biogenic atmospheres, machine learning

DOI: 10.47456/Cad.Astro.v5n1.43184

1 Introdução

Desde a descoberta do primeiro exoplaneta a orbitar uma estrela da Sequência Principal [1], um enorme esforço vem sendo realizado pela comunidade científica para tentar entender algumas questões básicas, relacionadas à existência destes novos mundos como, por exemplo, como são constituídos?; qual a constituição de suas atmosferas?; apresentam condições mínimas para serem habitáveis? etc. Metodologias, modelos e mis-

sões espaciais vêm sendo propostas e executadas há décadas, com a finalidade de compreender as características desta vasta população de planetas extrassolares descobertos até o momento.

Muitos destes esforços permitiram uma razoável compreensão e popularização de conceitos importantes como o de Zona Habitável de um sistema planetário [2]. Hoje, sabe-se que essa é a região em torno da estrela de um sistema extrassolar, onde há grandes possibilidades da existência de água no estado líquido. Essa informação é relevante, pois é sabido que a vida na Terra utilizou esse solvente para se desenvolver e evoluir. Tal constatação sempre foi utilizada para que planetas localizados nessa região fossem alvos prioritários para missões em busca de bioassinaturas, que pudessem indicar a presença de vida fora da Terra. Porém, planetas podem estar localizados no interior de Zonas Habitáveis e não possuírem condições para abrigar vida ou até mesmo reter uma atmosfera que possa ser analisada remotamente, que é a única maneira de se obter informações sobre a história biológica destes corpos.

Apesar de vários planetas extrassolares terem sido descobertos e estudados, pouco se conhece a respeito de sua constituição interna, o que impede uma análise mais rigorosa sobre as reais condições de sua superfície e de sua atmosfera. Heller e Armstrong [3] afirmam que um planeta mais massivo do que a Terra deve possuir uma superfície maior, propiciando uma quantidade de biomassa e biodiversidade mais elevada. Esse fato favorece a manutenção de uma atmosfera densa (e, provavelmente, contaminada com marcadores biológicos), já que a gravidade do exoplaneta é maior. Porém, pesquisas recentes indicam que planetas com duas massas terrestres são em sua maioria mininetunos, provavelmente com atmosferas ricas em gases leves como o hidrogênio, o que não favorece cenários similares aos encontrados na Terra [4].

Outros estudos apontam que planetas mais massivos poderiam ter o regime de movimentação de placas tectônicas comprometido e, consequentemente, ter o ciclo silicato-carbono não ativo [5,6] ou até mesmo evoluírem para planetas do tipo Vênus, com maiores atividades vulcânicas e emissão de gases [7,8].

Como vários aspectos desses novos mundos ainda não são conhecidos e muitos nem serão citados aqui, esta pesquisa analisou planetas com raios entre 0,3 – 2,5 raios terrestres e massas variando entre 0,01 – 10 massas terrestres, o que abrange planetas semelhantes a Marte (0,4 – 0,8 raios terrestres), à Terra (0,8 – 1,6 raios terrestres), a superterras (1.6 – 2.5 raios terrestres) e a Mininetunos (1,6 – 2,5 raios terrestres), orbitando estrelas do tipo espectral M (2400 – 3700 K), K (3700 – 5200 K) e G (5200 – 6000 K). Essa escolha foi realizada, pois há grandes chances de vários exoplanetas, contidos nessa faixa de análise, abrigarem atmosferas detectáveis. Ressalta-se, ainda, que as características dos exoplanetas (massa, raio etc.) e de seus sistemas (tipo espectral da estrela, luminosidade etc.) foram analisados com base em dados públicos disponibilizados pela National Aeronautics and Space Administration [9]. Foram estudados 72 sistemas planetários com o objetivo de se utilizar modelagens e técnicas de *machine learning* para identificar os exoplanetas mais aptos a possuírem atmosferas biogênicas. Na tentativa de analisar as relações entre sistemas planetários e seus planetas em uma perspectiva de clusterização e de análise multinível, foi utilizado o *Earth similarity index* (ESI), um índice de classificação de exoplanetas por similaridade com a Terra que varia entre 0 e 1 (quanto mais próximo de um mais semelhante à Terra). Vale lembrar que o índice não denota que o planeta seja realmente habitável [10].

Analisando o ESI quando aplicado no estudo de corpos celestes presentes no Sistema Solar, percebe-se que seus valores podem ser úteis para um estudo comparativo da similaridade deles com os exoplanetas descobertos. Marte, por exemplo, possui um valor de ESI global de 0,70, já Vênus um valor de 0,44, enquanto Júpiter 0,29, Saturno 0,25 e Netuno 0,18. Se forem consideradas as luas, alvos de estudos astrobiológicos, encontram-se Europa com ESI igual a 0,26, Titã com valor igual a 0,24, Enceladus com 0,094 e Io com 0,36 [10].

Com as características dos sistemas planetários, dos exoplanetas e o ESI, é possível realizar processos de clusterização e análises multiníveis com o objetivo de entender a correlação entre características dos planetas e de seus sistemas na configuração de potenciais alvos na busca de sinais biológicos, utilizando a Terra como modelo, mas considerando situações mais abran-

Tabela 1: Variáveis selecionadas. O período orbital, P_p, é dado em dias e a luminosidade estelar, L_s, é dada em termos do logaritmo da luminosidade solar. Fonte: dados originais de pesquisa oriundos do banco de dados da NASA [9].

Nome do Planeta	Planeta	Sistema	ESI	R_p	M_p	P_p	L_s
GJ 357 b	1	1	0,36	1,20	1,84	3,93	-1,80
GJ 357 c	2	1	0,43	1,66	3,40	9,12	-1,80
GJ 357 d	3	1	0,45	2,34 6,10	55,66	-1,80	
GJ 9827 b	4	2	0,22	1,58	5,14	1,21	-0,94
GJ 9827 c	5	2	0,27	1,24	1,24	3,65	-0,94
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
YZ Cet d	240	72	0,66	1,03	1,09	4,66	-2,66

gentes como, por exemplo, a análise de planetas extrassolares fora da zona habitável. Assim, entende-se que planetas mais semelhantes à Terra possuem uma probabilidade maior de abrigar uma atmosfera que agregue importantes informações sobre a sua evolução.

2 Materiais e métodos

Os dados utilizados na pesquisa se referem a algumas características de 240 exoplanetas, aninhados em 72 sistemas planetários e que podem ser observadas, em parte, na Tabela 1. Nela, são apresentados o nome do exoplaneta e a sua identificação de nível (1), assim como a identificação de nível (2) para os sistemas planetários. Ainda foram dispostos o valor do ESI, do raio do planeta, Rp (em raios terrestres), da massa do exoplaneta, Mp (em massas terrestres), do período orbital do planeta, Pp (dado em dias) e da luminosidade estelar, Ls (log da luminosidade solar), que é uma característica associada à estrela de cada sistema planetário.

Uma avaliação preliminar dos dados contidos na Tabela 1 aponta para a possibilidade da utilização da modelagem multinível nesta pesquisa. A estrutura básica dessa abordagem considera que os exoplanetas estão aninhados nos sistemas planetários, o que leva à criação de um modelo com dois níveis. A estruturação dos dados em níveis é vantajosa quando comparada àquelas que se utilizam de modelos clássicos de regressão linear, pois permite discutir determinadas correlações a partir de uma perspectiva hierárquica, considerando dependências entre observações pertencentes a um mesmo grupo, o que torna possível a captura de comportamentos de variáveis estudadas em cada um dos níveis propostos [11]. A hierarquia proposta aqui considera os exoplanetas constituindo o nível 1 e os sistemas planetários o nível 2 (vide Figura 1). Os dados referentes às características dos 240 exoplanetas abordados nesta pesquisa foram agrupados por semelhança utilizando o método não hierárquico k-means, que consiste no agrupamento de elementos amostrais em determinado *cluster* cujo centróide (vetor de média amostral) é o mais próximo do vetor de valores observados para o respectivo elemento [13]. O objetivo da utilização da técnica está relacionado à investigação das características dos planetas, que constituem as aglomerações e que podem estar associadas ao ESI. Como a técnica exige que se escolha previamente a quantidade de *clusters*, foram utilizados o método hierárquico aglomerativo e o método de *elbow* para definir esse valor [11].

Segundo Fávero e Belfiore [11], o método hierárquico Aglomerativo pode ser utilizado em casos em que todas as observações são consideradas separadamente e, a partir de suas distâncias ou semelhanças, sejam formados grupos até que se estabeleça um estágio final com apenas um agrupamento. Dentre os métodos existentes (*single linkage, complete linkage e average linkage*), optou-se pelo denominado encadeamento completo (*complete linkage*), que dá preferência às maiores distâncias entre as observações ou grupos na formação de novos agrupamentos. Ele é utilizado em situações em que não há afastamentos consideráveis entre as observações e que o objetivo final da investigação seja a identificação das heterogeneidades existentes entre elas. John-

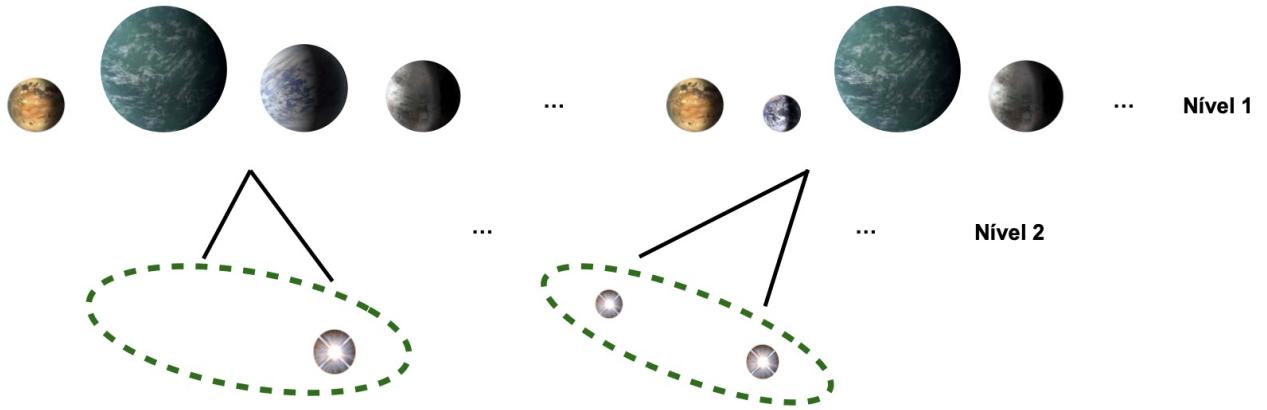

Figura 1: Esquema da Modelagem Multinível. Nível 1: exoplanetas; nível 2: sistemas planetários. Fonte: resultados originais de pesquisa [12].

son e Wichern [14] apresentam uma sequência de procedimentos que facilita a compreensão do método hierárquico aglomerativo, cujo resultado é um gráfico de árvore, denominado dendrograma, que explicita o processo de aglomeração, apontando a alocação de cada observação em cada agrupamento. A aplicação do método nos dados desta pesquisa gerou um dendrograma, sugerindo a existência de quatro *clusters*.

Outra técnica utilizada para a escolha da quantidade de *clusters* a ser inserida como *imput* no algoritmo *k-means* foi o método do cotovelo (*elbow method*). Ela é responsável por simular diversas divisões em número crescente de grupos e calcular as variâncias internas de cada um deles, almejando um ponto de equilíbrio [15]. Ao final do processo é gerado um gráfico a partir do qual é possível extrair a quantidade ideal de *clusters* a ser considerada. Para o conjunto de dados deste estudo, o método apontou quatro *clusters*, confirmando a estimativa oriunda do método *complete linkage*.

3 Clusterização

O estudo dos agrupamentos de exoplanetas pode ser realizado de forma mais consistente com a utilização da técnica de análise das componentes principais ou PCA (*principal component analysis*). A PCA é uma técnica matemática que é capaz de construir novas métricas ou variáveis, que passam a ser uma combinação linear das métricas originais. Segundo Dangeti (2017), ela pode ser utilizada como mecanismo de redução de dimensionalidade de um determinado conjunto de dados, sendo amplamente utilizada em etapas de pré-processamento, que visam o uso de métodos de clusterização como, por exemplo, o método *k-means*, permitindo uma melhor visualização dos dados em um espaço de duas dimensões, assim como a melhora da generalização dos algoritmos.

Com a finalidade de agrupar variáveis que apresentam correlações significativas entre si, foi realizada uma redução da dimensionalidade da base de dados, que no caso desse estudo apresenta quatro variáveis fundamentais: o raio e a massa do planeta, o ESI e a luminosidade estelar. A redução da dimensionalidade foi realizada por meio da técnica de análise de componentes principais (PCA), isto é, passou-se de quatro para duas dimensões (componente 1 e componente 2). A primeira componente representa 59,2% das informações contidas nas variáveis originais, enquanto a segunda, 29,7%, totalizando aproximadamente 88,8%. Assim, é possível trabalhar com duas dimensões, PC1 e PC2 (espaço bidimensional), que carregam consigo grande parte das informações contidas nas variáveis originais. Vale lembrar, que os fatores resultantes desta técnica são ortogonais entre si e, consequentemente, possuem correlação igual a zero e podem ser utilizados em situações que demandem ausência de colinearidade [11]. Dessa forma, os exoplanetas foram agrupados em quatro *clusters*, o que facilitou a compreensão das relações existentes entre eles.

A análise de PCA foi utilizada já que foram observadas correlações entre as variáveis de estudo. Como já esperado, há uma correlação positiva (0,84) entre a massa e o raio do exoplaneta,

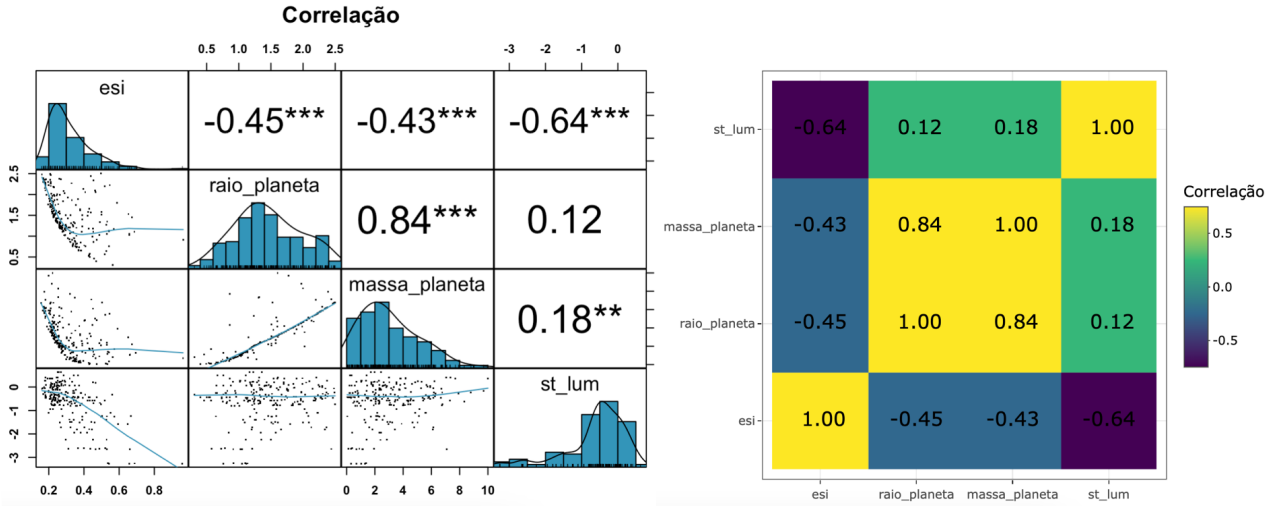

Figura 2: Correlações entre as variáveis de estudo e a relação entre ESI e luminosidade estelar; ESI é o índice de similaridade com a Terra, st_lum é a luminosidade estelar, raio_planeta é o raio do planeta e massa_planeta é a massa do planeta. Fonte: dados originais de pesquisa [12].

Tabela 2: Teste de esfericidade de Bartlett. Fonte: resultados originais de pesquisa [12]

	X^2	P-valor	Graus de liberdade
Esfericidade de Bartlett	496,76	0,00	6

Tabela 3: Autovalores e variância compartilhada. Fonte: resultados originais de pesquisa [12]

	PC1	PC2	PC3	PC4
Autovalor	2,36	1,18	0,30	0,14
Variância Compartilhada	59,20 %	29,70 %	7,50 %	3,60 %
Variância Cumulativa	59,20 %	88,80 %	96,40 %	100,00 %

confirmando que planetas mais massivos possuem raios maiores. Ficou evidenciado que a luminosidade estelar, característica da estrela do sistema planetário que pode ser entendida como a energia emitida por ela por unidade de tempo, possui uma correlação negativa ($-0{,}64$) com o ESI, mostrando a importância dessa variável de contexto na busca por planetas que possam abrigar atmosferas detectáveis. A Figura 2 mostra as correlações entre as variáveis de estudo.

A garantia de que os dados se adequavam à análise proposta ocorreu por meio do teste de esfericidade de Bartlett (vide Tabela 2), onde fica evidenciado que para um nível de significância de 5%, pode se rejeitar a hipótese nula de que a matriz de correlações de Pearson para as variáveis de estudo seja igual a uma matriz identidade de mesma ordem, ou seja, de que todas as correlações entre as variáveis apresentadas na Figura 2 sejam estatisticamente iguais a zero [11].

A análise dos fatores obtidos e a variância compartilhada capturada por cada um deles é mostrada na Tabela 3.

A escolha das duas primeiras componentes e a consequente redução de dimensionalidade ficam justificadas pelo critério de Kaiser, que considera a tomada de fatores com autovalor superior a um.

Segundo Fávero e Belfiore [11], o método de aglomeração não hierárquico *k-means* pode ser utilizado para dividir as observações contidas em um banco de dados em k *clusters*, garantindo que estejam mais próximas entre si quando comparada a qualquer outra pertencente a um diferente, ou seja, espera-se que a variabilidade entre os agrupamentos formados seja grande, mas que a variabilidade dentro deles seja pequena. O proce-

dimento requer que a quantidade de *clusters* seja definida a priori e, nesta pesquisa, optou-se por utilizar o método hierárquico aglomerativo (distância euclidiana / método *complete linkage*), que indicou a formação de quatro *clusters*, em conjunto com uso do método *elbow* (ou método cotovelo), que também sugeriu a formação de quatro *clusters*.

As variáveis escolhidas para clusterização foram: os índices ESI, os raios e as massas dos exoplanetas, além das luminosidades estelares. A escolha é uma forma de garantir que a clusterização terá como base ao menos duas características associadas à classificação dos exoplanetas (tipo Marte, Terra, superterras ou mininetunos) e ao menos uma correlacionada ao sistema planetário (luminosidade estelar), além do índice de similaridade com a Terra.

Ressalta-se que o objetivo da clusterização foi verificar como a distribuição dos exoplanetas pode gerar agrupamentos, que contenham informações relevantes sobre a similaridade deles com a Terra. A ideia é de que ESIs altos estejam associados a planetas mais propensos a abrigar atmosferas biogênicas. Foi estipulado um valor de corte em que exoplanetas com valores de ESI igual ou maior a 0,5 seriam aqueles com as maiores probabilidades de possuírem um conteúdo atmosférico. Como não há muitas informações reais sobre os envoltórios gasosos dos planetas extrassolares, o valor adotado pode ser utilizado para execução de algumas investigações, mas podendo ser revisto posteriormente.

A Figura 3 mostra os *clusters* formados a partir da utilização do método *k-means* e das componentes um e dois oriundas da redução de dimensionalidade proporcionada pela utilização da PCA. Vale ressaltar que os círculos azuis representam os centróides dos *clusters*, ou seja, pontos a partir dos quais as observações semelhantes se aglomeram.

Outras correlações que se apresentam passíveis de investigação são aquelas entre o raio do exoplaneta e o ESI e entre o período orbital e o ESI, já que se pretende estudar os mais diversos cenários. A Figura 4 mostra a distribuição dos valores do índice ESI em função dos raios, dos períodos orbitais e das massas dos exoplanetas, onde fica evidenciado que valores de similaridade acima de 0,50 estão distribuídos em uma ampla faixa de valores de raios e massas. Quando se observa a relação entre os períodos orbitais e o ESI, percebe-se que valores acima 0,5 possuem, também, a tendência de se distribuírem por uma faixa diversificada de valores, porém temos um viés que está correlacionado com a quantidade de dados disponíveis para análise (poucos planetas com períodos orbitais maiores detectados). Desse modo, não é prudente a desconsideração de nenhum tipo de exoplaneta quando se pensa na busca por atmosferas biogênicas extraterrestres.

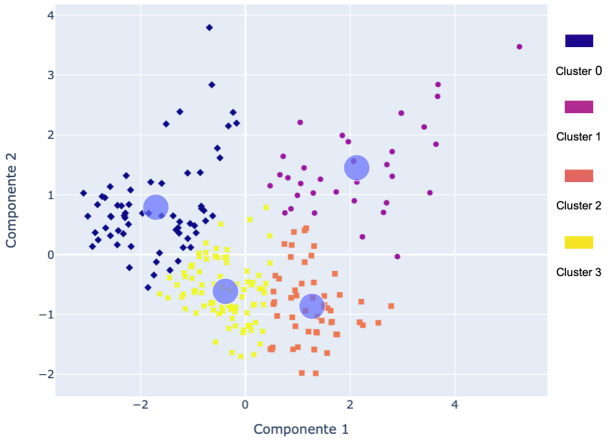

Figura 3: Distribuição dos planetas por *cluster*. Os círculos azuis representam os centros de aglomeração. Fonte: resultados originais de pesquisa [12].

A análise da Figura 4 e dos dados contidos na Tabela 4, referentes aos *clusters* formados e apresentados na Figura 3, mostram que o conjunto de exoplanetas que apresenta os maiores valores para a variável ESI, possui raios variando entre 0,57 e 1,80 raios terrestres e massas, com valores entre 0,07 e 3,89 massas terrestres, constituindo o *cluster* número 1. Também é perceptível que os *clusters* 0 e 2 possuem uma população de planetas com valores para o ESI médio inferiores quando comparado à média do *cluster* 1, porém abriga exoplanetas com valores para o ESI acima de 0,50. Já o *cluster* 3, possui objetos com baixos valores para o ESI, sendo que o maior valor é igual 0,42. Entretanto, percebe-se que todos os *clusters* formados comportam planetas semelhantes a Marte, Terra, superterras e mininetunos, o que aponta para uma considerável diversidade de exoplanetas com propensão para abrigar atmosferas e passíveis de serem estudadas remotamente.

É possível apontar os exoplanetas que possuem a maior probabilidade de abrigar envoltórios ga-

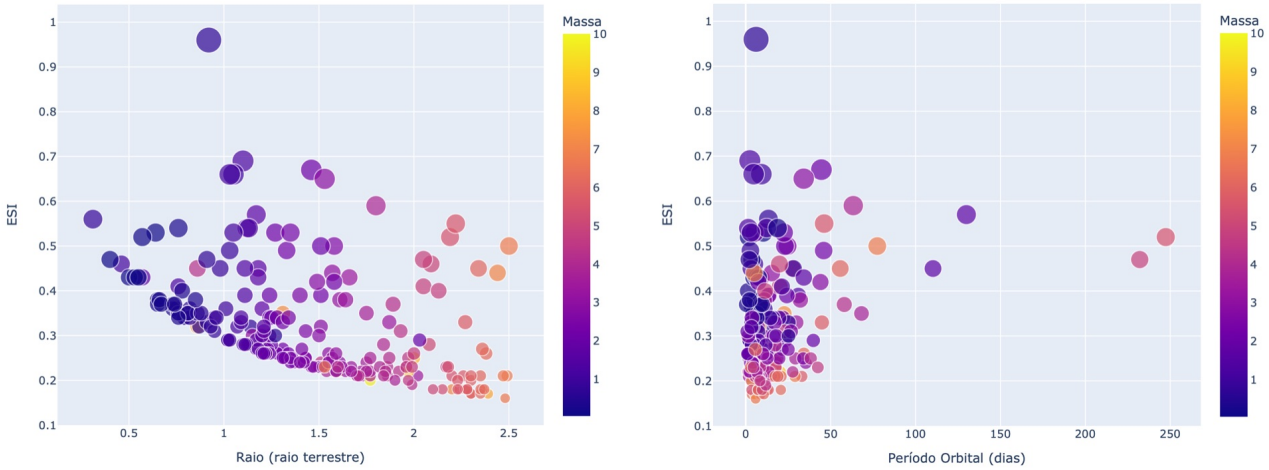

Figura 4: Relação entre ESI e raio e entre ESI e período orbital de exoplanetas. A massa na escala de cores é apresentada em unidades terrestres. Fonte: dados originais de pesquisa [12].

Tabela 4: Características dos *clusters* formados pelo algoritmo *k-means*. Fonte: resultados originais de pesquisa [12]

Dados estatísticos	*cluster* 0				*cluster* 1			
	ESI	Rp	Mp	Ls	ESI	Rp	Mp	Ls
Média	0,25	2,09	5,69	– 0,50	0,50	1,21	1,72	– 2,16
Desvio Padrão	0,09	0,27	1,34	0,58	0,13	0,32	1,04	0,72
Valor mínimo	0,16	1,31	3,53	– 2,64	0,30	0,57	0,07	– 3,26
25 %	0,19	1,96	4,66	– 0,73	0,39	1,03	0,94	– 2,66
50 %	0,22	2,15	5,60	– 0,46	0,50	1,19	1,76	– 2,10
75 %	0,26	2,30	6,21	– 0,06	0,54	1,51	2,41	– 1,50
Valor máximo	0,55	2,50	10,00	0,41	0,96	1,80	3,89	– 1,23
Nº de planetas	65				34			

Dados estatísticos	*cluster* 2				*cluster* 3			
	ESI	Rp	Mp	Ls	ESI	Rp	Mp	Ls
Média	0,36	0,90	0,99	– 0,44	0,26	1,44	2,82	– 0,17
Desvio Padrão	0,06	0,25	0,76	0,46	0,04	0,22	0,94	0,40
Valor mínimo	0,27	0,31	0,04	– 1,71	0,20	0,86	1,37	– 1,50
25 %	0,33	0,76	0,44	– 0,63	0,23	1,24	2,18	– 0,50
50 %	0,35	0,89	0,80	– 0,43	0,25	1,43	2,63	– 0,12
75 % 0,39	1,12	1,40	– 0,24	0,27	1,60	3,33	0,09	
Valor máximo	0,56	1,33	4,30	0,50	0,42	2,03	7,70	0,63
Nº de planetas	56				85			

sosos, capazes de conter materiais biológicos passíveis de serem observados, considerando o índice de similaridade com a Terra. A Tabela 5 apresenta aqueles planetas que possuem valores para o ESI maiores ou iguais a 0,5 e seus respectivos *clusters*.

Os resultados oriundos da aplicação das técnicas de *machine learning*, demonstram que há uma variedade grande de exoplanetas que podem servir de alvos na busca por bioassinaturas. É evidente que alguns se mostram mais propensos a possuir uma atmosfera biogênica, pois pos-

Tabela 5: Exoplanetas com maior propensão para abrigar atmosferas biogênicas. Fonte: resultados originais de pesquisa [12]

Planeta	ESI	Sistema Planetário	*cluster*
K2-3c	0,50	6	3
K2-3d	0,67	6	1
Kepler 138-b	0,53	15	1
Kepler 138-d	0,50	15	3
Kepler 174-d	0,52	22	0
Kepler 186-e	0,53	24	1
Kepler 186-f	0,57	24	1
Kepler 235-e	0,55	33	0
Kepler 296-e	0,65	41	1
Kepler 296-f	0,59	41	1
Kepler 298-d	0,50	42	1
Kepler 37-b	0,56	54	2
Kepler 42-d	0,52	58	1
Kepler 446-d	0,53	62	1
Trappist-1b	0,54	71	1
Trappist-1c	0,69	71	1
Trappist-1e	0,96	71	1
Trappist-1f	0,66	71	1
Trappist-1g	0,54	71	1
Trappist-1h	0,54	71	1
YZ Cet c	0,53	72	1
YZ Cet d	0,66	72	1

suem um grau de similaridade com a Terra maior. Percebe-se, também, que dos 72 sistemas extrassolares estudados, apenas 12 possuem exoplanetas com ESI maior ou igual a 0,5.

Ressalta-se que a pesquisa tem como proposta entender o motivo pelo qual a similaridade com a Terra é mais elevada para certos grupos de planetas. Será que ela se deve somente às condições intrínsecas dos exoplanetas ou está fortemente correlacionada às características do sistema planetário onde eles estão localizados? Dessa forma, o objetivo do trabalho não é apenas apresentar uma equação que forneça uma estimativa do grau de similaridade de exoplanetas com a Terra e, consequentemente, expor aqueles mais aptos a apresentarem um envelope gasoso, mas também, de tentar fornecer a porcentagem de contribuição dos fatores ligados às características dos planetas e das particularidades dos sistemas planetários na constituição desse índice de similaridade, com o intuito de quantificar essa informação.

4 Modelagem multinível HLM2 nulo (modelos hierárquicos lineares de dois níveis)

A tentativa de entender como a variável ESI pode estar correlacionada com parâmetros dos exoplanetas e dos sistemas planetários onde estão localizados passa, necessariamente, pela aplicação de uma modelagem multinível. Assim, considerou-se que os exoplanetas (nível 1) estão aninhados a sistemas planetários (nível 2), conforme mostra a Figura 1. Os modelos multiníveis incorporam naturalmente a estrutura de agrupamento dos objetos a serem estudados e podem ser bem complexos. Neste trabalho, apesar do estudo da decomposição da variância nos diferentes níveis de análise do fenômeno físico de interesse, não foram incluídas variáveis explicativas. Segundo Fávero e Belfiore [11], essa inclusão agregaria investigações adicionais das relações entre essas variáveis e a variável dependente, já que seriam consideradas interações atreladas aos distintos níveis nos componentes de efeitos fixos e aleatórios, o que tornaria o modelo mais sofisticado e apto a investigar outros cenários.

O propósito não é somente verificar se há variabilidade no valor do ESI, quando se considera planetas que estão localizados em sistemas planetários distintos, mas sim quantificar esse efeito. Seguindo Fávero e Belfiore [11], a verificação pode ser iniciada com a utilização de um modelo multinível nulo (HLM2-nulo), já que nesse tipo de modelagem não há procedimentos do tipo *stepwise*. Desse modo, utilizou-se a metodologia baseada no *step-up strategy* [16], que consiste em criar modelos cada vez mais complexos, a partir da inserção de efeitos aleatórios de interceptos, de inclinação, assim como de variáveis explicativas e contextuais, comparando suas eficiências por meio, por exemplo, do ganho de *log-likelihood* (teste de verossimilhança).

Aqui, é apresentado somente o resultado da análise do comportamento de um modelo multinível simplificado HLM2 nulo, ou seja, a primeira etapa do método *step-up strategy*. O modelo proposto pode ser descrito pelas seguintes equações

$$\text{Nível 1} \longrightarrow \text{ESI}_{ij} = \beta_{0j} + \varepsilon_{ij}, \qquad (1)$$

e

$$\text{Nível 2} \longrightarrow \beta_{0j} = \gamma_{00} + \nu_{0j}, \qquad (2)$$

Tabela 6: Modelo HLM2 nulo. $var(\nu_{0j})$ e $var(e)$ são as variâncias dos termos de erro e são utilizadas para calcular a correlação intraclasses (ICC). Fonte: resultados originais de pesquisa [12]

Componentes	Variância	Desv. Pad.	P-valor
$var(\nu_{0j})$	0,007	0,001	0,000
$var(e)$	0,005	0,001	0,000

Coeficientes	Estimativa	t valor	Desv. Pad	P-valor
Intercepto	0,309	27,887	0,011	0,000

ou seja,

$$\mathrm{ESI}_{ij} = \gamma_{00} + \nu_{0j} + \varepsilon_{ij}, \tag{3}$$

onde γ_{00} é o valor do intercepto (representa a média global do ESI), ν_{0j} o efeito aleatório de intercepto (associado ao nível 2 se relacionando com o afastamento do ESI médio do planeta j à média global) e ε_{ij} o termo de erro (efeito aleatório associado ao nível 1 – indivíduo ij).

O modelo descrito pela equação 3 não considera variáveis explicativas e, mesmo assim, apresenta resultados interessantes, principalmente quando comparado ao modelo de mínimos quadrados ordinários (MQO) no que se refere a valores de *log-likelihood*, que apresenta um ganho substancial a partir da inclusão de um grau de liberdade a mais (214,37 contra 172,584). O resultado da análise do HLM2 nulo é apresentado na Tabela 6.

O modelo HLM2 nulo já é capaz de mostrar que existe variabilidade no valor do ESI entre exoplanetas provenientes de sistemas planetários distintos, mesmo que não tenham sido incorporadas nele variáveis explicativas ou de contexto como, por exemplo, o período orbital, a temperatura de equilíbrio, luminosidade estelar etc., ou seja, apenas considerando a existência de um intercepto γ_{00} e dos termos de erro ν_{0j} e ε_{ij}. Esse efeito ocorre, pois o fato de ν_{0j} ser diferente de zero indica que as "curvas" associadas a cada sistema planetário não possuem γ_{00} iguais.

Assim, as diferenças entre os valores do ESI para planetas que estão localizados em sistemas extrassolares distintos são justificadas, pois a variância do termo ν_{0j} é estatisticamente significante. Essa constatação pode parecer óbvia, mas o que se busca é uma quantificação do efeito que o ambiente planetário exerce na constituição do valor do ESI e, consequentemente, na probabilidade de certo planeta abrigar um conteúdo gasoso. A importância relativa do efeito sistema planetário sobre o valor do ESI pode ser estimada calculando a correlação intraclasses (ICC), que é a razão entre o efeito sistema planetário e a soma das variâncias. Com o cálculo do ICC, $\mathrm{ICC} = \frac{0{,}007}{0{,}007+0{,}006}$, que apontou um valor aproximado para este modelo simplificado de 0,54, ficou evidenciado que 54% da variação do valor do ESI é devido ao efeito sistema planetário.

A equação que descreve o modelo HLM2 nulo pode ser reescrita como

$$\mathrm{ESI}_{ij} = 0,309 + \nu_{0j} + \varepsilon_{ij} \tag{4}$$

Seguindo Tabachnick e Fidell [17] e observando que a variância do termo ν_{0j} se mostrou significante, percebe-se que não é prudente a utilização de um modelo de regressão linear tradicional (OLS - *ordinary least squares*) para o estudo do fenômeno de interesse, que está ligado a análises de contextos complexos como a interação planeta - sistema planetário. Ou seja, a relação entre o ESI e outras variáveis como, raio, massa, luminosidade etc., deve ser estudada considerando modelagens multiníveis.

5 Conclusão

Por meio das análises estatísticas desenvolvidas nesta pesquisa, é possível concluir, mesmo considerando certas limitações, que há uma grande variedade de exoplanetas capazes de abrigar atmosferas biogênicas. Essa conclusão é baseada no índice ESI e na noção de que quanto mais parecido com a Terra, maiores são as chances de eles possuírem um envoltório gasoso detectável. A limitação ocorre, pois há poucas informações reais sobre a constituição interna desses corpos planetários, além do fato de que foi estipulado um valor de corte para o ESI (maior ou igual a 0,5)

que está associado à presença de envelopes gasosos, podendo até mesmo ser considerado otimista ou mesmo pessimista, já que não existem muitas informações sobre as atmosferas dos exoplanetas descobertos.

A partir da análise realizada com o modelo não hierárquico *k-means*, foi possível indicar alguns exoplanetas que apresentam maiores probabilidades de possuírem uma atmosfera, contendo, porventura, algum material biológico passível de detecção remota. O modelo foi capaz de agrupar em *clusters* exoplanetas, que possuem entre si e com a Terra algum grau de semelhança, deixando evidente que o *cluster* 1 é aquele que comporta os principais candidatos. Assim, planetas como Trappist-1e, Trappist-1f, Trappist-1c, K2-3d, Kepler 296-e, YZ Cet d, entre outros, devem ser alvos prioritários na busca por bioassinaturas, pois possuem ESI acima do valor utilizado como limite. É interessante observar que esses planetas orbitam estrelas com luminosidades dentro de uma faixa específica, que vai de −3,26 a −1,23 (log da luminosidade solar); percebe-se, também, que eles possuem raios entre 0,57 e 1,80 raios terrestres e massas entre 0,07 e 3,89 massas terrestres. É um *cluster* que pode abrigar planetas do tipo Marte, Terra, superterra e mininetunos, mostrando a diversidade de objetos passíveis de estudo.

A análise multinível indicou que os sistemas planetários desempenham um papel importante na evolução de seus exoplanetas, o que influencia a constituição do ESI, mostrando que efeitos como a luminosidade estelar, a temperatura estelar, a metalicidade estelar, entre outras características não estudadas neste trabalho, podem ser fundamentais no estudo do grau de similaridade desses planetas com a Terra. Ficou evidenciado de modo quantitativo, a partir do modelo multinível HLM2 nulo e dos dados relativos aos 240 exoplanetas estudados, que mais de 50% da variação do ESI é oriunda das características dos sistemas planetários.

Fica evidente que modelagens multiníveis podem ser ótimas ferramentas para a compreensão de sistemas complexos como aqueles relacionados à habitabilidade planetária, já que os modelos podem ser escalados, considerando níveis de complexidade maiores como, por exemplo, a inserção de variáveis explicativas e contextuais, assim como a introdução da evolução temporal dos ambientes estudados.

Todos os códigos utilizados nesta pesquisa podem ser consultados por meio do link: https://github.com/luander22/Exoplanetas.git. Eles são oriundos da Ref. [11].

Sobre os autores

Natural de Ribeirão Preto - SP, Luander Bernardes (luander.uspicio@gmail.com) é Bacharel em Física Médica pela Universidade de São Paulo (USP) e Licenciado em Física pelo Instituto Federal de Educação, Ciência e Tecnologia de São Paulo. Possui Especialização em Ensino de Física pela Universidade Cruzeiro do Sul e em Data Science e Analytics pela USP. É Mestre e Doutor em Astrofísica pelo Instituto de Astronomia, Geofísica e Ciências Atmosféricas (IAG/USP). Atualmente é professor substituto no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, docente dos cursos presenciais e EAD do Centro Universitário Estácio de Ribeirão Preto (Ciências Exatas e Licenciaturas em Física e Química), além de também coordenar estes cursos na modalidade EAD. Possui experiência nas áreas de astrobiologia, espectroscopia no infravermelho, modelos atmosféricos, habitabilidade planetária e técnicas de *machine learning* aplicadas ao estudo de exoplanetas.

Anna Carolina Martins (anna.martins@gmail.com) é natural de Montes Claros - MG. Bacharel em Economia pela Universidade Federal de Campinas, Mestra em Economia pela Universidade Federal de Ouro Preto, Doutora em Economia pela Universidade Estadual de Montes Claros e Especialista em Controladoria e Finanças pela USP. Atuou como pesquisadora visitante na Universidade de Cambridge e na Universidade de Sant'anna (2022-2023) pelo projeto *Economics of energy innovation and system transition* (EEIST). Possui experiência na área de teoria econômica, economia computacional, economia aplicada, modelos econométricos, mercado de energia e transição energética.

Referências

[1] M. Mayor e D. Queloz, *A Jupiter-mass companion to a solar-type star*, Nature **378**(6555), 355 (1995).

[2] J. F. Kasting, D. P. Whitmire e R. T. Reynolds, *Habitable Zones around Main Sequence Stars*, Icarus **101**(1), 108 (1993).

[3] R. Heller e J. Armstrong, *Superhabitable Worlds*, Astrobiology **14**(1), 50 (2014).

[4] L. Zeng, D. Sasselov e S. Jacobsen, *Mass-Radius Relation for Rocky Planets based on PREM*, The Astrophysical Journal **819**, 127 (2016).

[5] V. Stamenković et al., *The influence of pressure-dependent viscosity on the thermal evolution of super-Earths*, The Astrophysical Journal **748**(1), 41 (2012).

[6] L. Noack e D. Breuer, *Plate tectonics on rocky exoplanets: Influence of initial conditions and mantle rheology*, Planetary and Space Science **98**, 41 (2014).

[7] C. Dorn, L. Noack e A. B. Rozel, *Outgassing on stagnant-lid super-Earths*, Astronomy & Astrophysics **614**, A18 (2018).

[8] L. Noack, A. Rivoldini e T. Van Hoolst, *Volcanism and outgassing of stagnant-lid planets: Implications for the habitable zone*, Physics of the Earth and Planetary Interiors **269**, 40 (2017).

[9] National Aeronautics and Space Administration (NASA), *NASA Exoplanet Archive*, Página da internet. Disponível em `https://exoplanetarchive.ipac.caltech.edu/index.html`, acesso em mar. 2024.

[10] D. Schulze-Makuch et al., *A Two-Tiered Approach to Assessing the Habitability of Exoplanets*, Astrobiology **11**(10), 1041 (2011).

[11] L. P. Fávero e P. Belfiore, *Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata* (Elsevier, Rio de Janeiro, 2017).

[12] L. Bernardes, *Machine Learning: uma ferramenta para o estudo de exoplanetas que podem apresentar atmosferas biogênicas*, Monografia (Especialização em Data Science e Analytics), Escola Superior de Agricultura Luiz de Queiroz – Universidade de São Paulo (2023).

[13] J. A. Hartigan e M. A. Wong, *Algorithm AS 136: A K-Means Clustering Algorithm*, Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100 (1979).

[14] R. A. Johnson e D. W. Wichern, *Applied multivariate statistical analysis* (Pearson Education, 2007), 6^a ed.

[15] P. Dangeti, *Statistics for Machine Learning* (Packt Publishing, 2017).

[16] S. W. Raudenbush e A. S. Bryk, *Hierarchical linear models: applications and data analysis methods* (Thousand Oaks Sage Publications, 2002).

[17] B. Tabachnick e L. Fidell, *Using Multivariate Statistics: Pearson New International Edition* (Editora Pearson, 2013).