BREVES NOTAS SOBRE ENTROPIA E SUAS APLICAÇÕES

Paulo Saraiva
CMUC - Centre for Mathematics of the University of Coimbra; CeBER - Centre for Business and Economics Research.
psaraiva@fe.uc.pt

1. INTRODUÇÃO

A teoria da informação é a área da matemática que estuda a quantificação, o armazenamento e a comunicação da informação digital. Na sua base está a ideia de quantificar a informação existente em eventos aleatórios. Trata-se de uma área fundamentalmente estabelecida após os trabalhos de H. Nyquist [4] e R. Hartley [1] na década de 1920, mas principalmente por Claude E. Shannon na década de 1940, no seu impactante artigo "A Mathematical Theory of Communication" [5] (razão pela qual este matemático é conhecido por *pai da Teoria da Informação*), situando-se na interseção de várias disciplinas: Teoria das Probabilidades, Estatística, Ciências da Computação, Mecânica Estatística, Engenharia da Informação e Engenharia Eletrotécnica. Aplicações de tópicos fundamentais da teoria da informação incluem a codificação de fontes de informação, a compressão de dados (*e.g.*, para arquivos ZIP), a codificação de canais, bem como a deteção e correção de erros. O conceito chave na teoria da informação é a entropia, a qual constitui uma métrica para o grau de casualidade ou de incerteza que eventos aleatórios possuem. Entropia e quantidade de informação relacionam-se do seguinte modo: quanto maior a quantidade de informação, maior será a desordem e maior será a entropia; quanto menor a quantidade de informação, menor será a escolha e menor a entropia.

No presente artigo, recorrendo a um exemplo, apresenta-se de um modo construtivo e intuitivo q.b. a noção de entropia, ao mesmo tempo que se explica em que sentido esta serve como métrica para a quantidade de informação subjacente a certos fenómenos. Dentre as propriedades da entropia, deduz-se a que diz respeito à entropia máxima. A entropia de Shannon veio a ser adaptada a outras áreas do conhecimento, sendo uma das ferramentas mais recorrentes na medição da diversidade biológica (leia-se, *e.g.*, [3], [6] e [2]), justificando-se assim a inclusão de um exemplo da sua aplicação neste contexto. Por último, a entropia pode igualmente ser adaptada no sentido de quantificar a diversidade de origens geográficas de populações que migram, como, por exemplo, os estudantes universitários. Apresenta-se um exemplo original desta aplicação a um universo restrito de estudantes da Universidade de Coimbra, o qual, se alargado a um universo mais amplo, pode revelar-se importante como auxílio à caracterização de migrações estudantis.

Apresenta-se aqui uma breve e intuitiva introdução à entropia de Shannon, incluindo algumas das suas propriedades. Ilustra-se a aplicação desta medida de informação em dois contextos distintos do que esteve na sua génese.

2. ENTROPIA DE SHANNON: CONSTRUÇÃO INTUITIVA, DEFINIÇÃO E PROPRIEDADES

Para entender os conceitos de entropia e de quantidade de informação, considere o seguinte exemplo. Suponha que duas máquinas, M_1 e M_2, geram sequências de letras a partir de um alfabeto com apenas quatro letras, digamos A, B, C e D. M_1 gera cada letra aleatoriamente, de tal modo que cada uma ocorre em média 25% das vezes, enquanto M_2 gera as letras de acordo com as seguintes probabilidades de ocorrência:

$$p(A) = 50\%,\ p(B) = p(C) = 12,5\% \text{ e } p(D) = 25\%.$$

Qual das máquinas está a produzir mais informação? Claude Shannon refez a questão, colocando-a do seguinte modo: se tivesse de prever que letra deveria aparecer a seguir na sequência produzida por cada máquina, qual seria, em cada caso, o número mínimo de questões binárias que teria de realizar? Por questão binária, entendemos uma questão que divida as possibilidades em duas (isto é, questões de "sim ou não").[1]

Em relação à máquina M_1, a nossa primeira questão poderia ser: a próxima letra pertencerá ao conjunto $\{A, B\}$? A probabilidade de pertencer ao conjunto destas duas letras é de 50% e o mesmo sucede com a probabilidade de pertencer a $\{C, D\}$. Após obtermos a resposta (sim/não pertence), podemos eliminar metade das possibilidades e ficaremos apenas com duas letras, ambas equiprováveis. Por exemplo, se a resposta fosse afirmativa, poderíamos

[1] Uma interessante simulação de ambas as situações pode ser consultada em [7].

eliminar $\{C, D\}$ e colocar agora a questão: a letra seguinte é A? Após esta segunda questão, teremos identificado corretamente a próxima letra. Portanto, podemos dizer que a incerteza da máquina M_1 é igual a dois (duas questões por cada letra a adivinhar), uma vez que, independentemente do símbolo considerado, teremos de realizar duas questões se quisermos ter a certeza da próxima letra. A seguinte árvore binária, contendo as probabilidades associadas a cada questão colocada, pretende ilustrar esta situação.

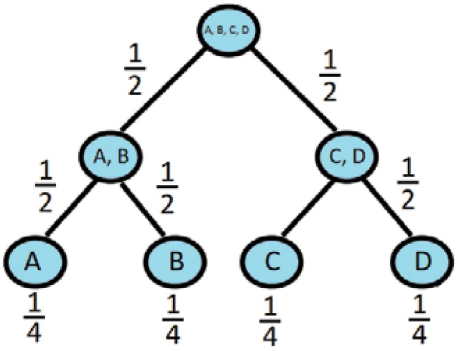

Figura 1. Árvore binária relativa a M_1.

Em relação à máquina M_2, à semelhança de M_1, bastariam duas questões para adivinhar a próxima letra. No entanto, o facto de as probabilidades de cada letra não serem todas iguais leva-nos a colocar as questões de maneira distinta. No caso presente, A ocorre com probabilidade 50%, sendo 50% a soma das probabilidades de ocorrência das restantes. Podemos começar por perguntar: será um A? No caso afirmativo, bastará uma pergunta. No caso negativo, ficámos com as restantes três letras: D, B e C, sendo estas últimas equiprováveis. Podemos fazer uma segunda pergunta: será um D? No caso afirmativo, bastaram-nos duas perguntas; no caso negativo, teremos de realizar uma terceira pergunta para identificar qual das restantes duas letras é a certa. Em média, quantas questões teremos de efetuar para identificar uma letra produzida pela máquina M_2? A árvore binária com as probabilidades associadas a cada questão colocada vai ajudar-nos a responder (figura 2).

Para calcular o número médio de questões que nos levam a cada uma das letras, tomaremos a média ponderada pelo número de questões:

$$p(A) \times 1 + p(D) \times 2 + p(B) \times 3 + p(C) \times 3 = 1{,}75.$$

Ou seja, no segundo caso, em média, o número expectável de questões binárias necessárias para atingir cada letra

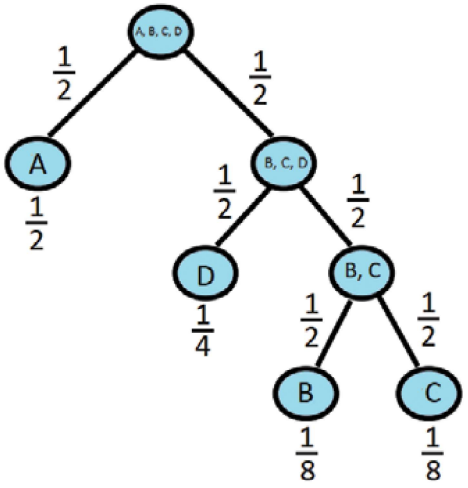

Figura 2. Árvore binária relativa a M_2.

é de 1,75, o que compara com as duas questões que, em média, necessitamos no caso da máquina M_1. Quer isto dizer que, se tivéssemos de adivinhar uma sequência de 100 símbolos gerados por cada uma das máquinas, seria expectável necessitarmos de 200 questões no primeiro caso, e de 175 questões no segundo caso. Isto significa que M_2 está a produzir menos informação do que M_1, uma vez que há menos incerteza ou surpresa acerca da letra a gerar. Claude Shannon chamou **entropia** a esta medida de incerteza média, tendo optado pela letra H para a representar. O **bit** (de *binary digit*), a unidade de informação escolhida por Shannon para H, baseia-se na incerteza inerente ao lançamento de uma moeda equilibrada, e equivale ao número médio de questões na analogia acima explanada.

Procuremos generalizar o conceito para o caso de n símbolos possíveis. Ainda segundo a analogia anterior, a entropia será o somatório, para cada símbolo, da sua probabilidade de ocorrência, p_i, multiplicada por s_i, número de questões binárias necessárias para o alcançar:

$$\mathrm{H} = \sum_{i=1}^{n} p_i \times s_i. \qquad (1)$$

A questão imediata é a de saber como representar s_i de uma maneira mais geral. Como pudemos observar, o número de questões depende do nível em que se encontra cada letra isolada no diagrama de árvore que modeliza cada máquina. Segundo esta analogia, cada símbolo isolado num determinado nível tem probabilidade $\frac{1}{2}$ relativamente ao nó que o originou. Assim, cada letra isolada no nível k tem probabilidade inicial $p = \frac{1}{2^k}$. Neste contexto, tem-se

$$p_i = \frac{1}{2^{s_i}},\ i = 1, \ldots, n,$$

e o número de questões para atingir o i-ésimo símbolo será

$$s_i = \log_2\left(\frac{1}{p_i}\right),\ i = 1, \ldots, n.$$

Substituindo em (1) esta última expressão, vem:

$$\mathrm{H}(p_1, \ldots, p_n) = \sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{n} p_i \log_2 p_i. \quad (2)$$

Esta é a chamada **entropia de Shannon** para um evento aleatório com n estados possíveis, com probabilidades p_i , $i = 1, ..., n$, onde $\sum_{i=1}^{n} p_i = 1$. Sempre que ocorra $p_i = 0$ para algum $i \in \{1, \ldots, n\}$, adotaremos a convenção

$$p_i \log_2 p_i = 0. \quad (3)$$

Note que os cálculos de H podem ser feitos diretamente a partir dos dados das frequências absolutas de cada estado. Sejam $f_1, \ldots, f_n$ tais frequências e S o número total de observações. Então:

$$\sum_{i=1}^{n} f_i = S \quad \text{e} \quad p_i = \frac{f_i}{S},\ i = 1, \ldots, n.$$

Deste modo,

$$\begin{aligned} \mathrm{H}(f_1, \ldots, f_n) &= -\sum_{i=1}^{n} p_i \log_2 p_i = -\sum_{i=1}^{n} \left(\frac{f_i}{S}\right) \log_2\left(\frac{f_i}{S}\right) \\ &= -\sum_{i=1}^{n} \frac{f_i}{S} \left[\log_2 f_i - \log_2 S\right] \\ &= \log_2 S - \tfrac{1}{S} \sum_{i=1}^{n} f_i \log_2 f_i. \end{aligned}$$

Admita que num dado evento aleatório existem n estados (ou categorias) possíveis, com probabilidades $p_i, i = 1, ..., n$, $n > 1$, tal que $\sum_{i=1}^{n} p_i = 1$. Designando por H_{max} a máxima entropia, vamos provar que

$$\mathrm{H}_{max} = \mathrm{H}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_2 n,$$

ou seja, que a quantidade de informação será máxima quando todos os estados são equiprováveis: $p_1 = \ldots = p_n = \frac{1}{n}$. Para tal, vamos recorrer a certas propriedades de $x \log_2 x$.

Consideremos a função F tal que

$$F(x) = \begin{cases} 0, \text{ se } x = 0 \\ x \log_2 x, \text{ se } x \in]0, 1] \end{cases} = \begin{cases} 0, \text{ se } x = 0 \\ kx \ln x, \text{ se } x \in]0, 1] \end{cases}, \quad (4)$$

onde $k = \frac{1}{\ln 2} > 0$. Após levantamento da indeterminação, mostra-se que

$$\lim_{x \to 0^+} F(x) = \lim_{x \to 0^+} kx \ln x = 0,$$

garantindo a continuidade de F em [0,1] (formalizando a convenção (3)). Note ainda que:

$$F'(x) = k \ln x + k \quad \text{e} \quad F''(x) = \frac{k}{x} > 0, \text{ para todo o } x \in]0, 1[.$$

Assim sendo, F é convexa (e o seu mínimo é $-ke^{-1}$, atingido no ponto $x = e^{-1}$). Recordemos agora um conhecido resultado das funções regulares.

Teorema 2.1 (Teorema do Valor Médio) *Seja f uma função contínua em [a,b] e diferenciável em]a,b[. Então existe algum $c \in]a, b[$ tal que:*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Para a demonstração que estamos a construir é importante destacar o seguinte resultado:

Corolário 2.1 *Seja f uma função contínua em [a,b] e diferenciável em]a,b[. Admita que f' é crescente em [a,b]. Para cada $t \in]a, b[$, se $p \in [a, b]$, então $f(p) \geq f(t) + f'(t)(p - t)$.*

Note que a função F definida em (4) satisfaz as condições do Teorema do Valor Médio, bem como do seu corolário, em[0,1]. Tomando $t = \frac{1}{n}$, o corolário permite concluir que, para cada $p_i \in [0, 1]$:

$$F(p_i) \geq F\left(\frac{1}{n}\right) + F'\left(\frac{1}{n}\right)\left(p_i - \frac{1}{n}\right).$$

Assim,

$$\begin{aligned} \sum_{i=1}^{n} p_i \log_2 p_i = \sum_{i=1}^{n} F(p_i) &\geq \sum_{i=1}^{n} \left[F\left(\frac{1}{n}\right) + F'\left(\frac{1}{n}\right)\left(p_i - \frac{1}{n}\right)\right] \\ &= nF\left(\tfrac{1}{n}\right) + F'\left(\tfrac{1}{n}\right)\left[\sum_{i=1}^{n} p_i - 1\right] \\ &= nF\left(\tfrac{1}{n}\right) = -\log_2 n. \end{aligned}$$

Logo,

$$\mathrm{H}(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \leq \log_2 n = \mathrm{H}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right),$$

concluindo a prova.

Sempre que uma situação traduza um afastamento da equiprobabilidade, ou caso se introduza algum grau de previsibilidade, a entropia deverá diminuir. Além disso, a entropia será nula quando a totalidade das ocorrências pertencer a um único estado. Fica assim clara a ideia de que, se a entropia de uma fonte de informação diminuir,

podemos fazer, em média, menos questões para adivinhar o resultado gerado. O número de bits dá-nos então uma medida quantitativa da informação, da surpresa ou da incerteza associada a uma determinada distribuição aleatória.

Admita que X é uma variável aleatória assumindo n estados possíveis, $x_1, \ldots, x_n$, e que p é a sua distribuição de probabilidade. Considere que as probabilidades de ocorrência de cada um dos n estados possíveis são dadas por $p_1,\ p_2, \ldots, p_n$ $(p_i = p(X = x_i))$. De acordo com Shannon, toda a função de entropia $\mathrm{H} = \mathrm{H}(p_1, \ldots, p_n)$ deverá assumir as seguintes propriedades:

1. A entropia é contínua enquanto função de cada $p_i,\ i = 1, \ldots, n$.

2. Se $p_i = \frac{1}{n}$, então a entropia é uma função monótona crescente de n . Para eventos equiprováveis, quanto maior for o número de acontecimentos possíveis, maiores a incerteza e a possibilidade de escolha.

3. Se uma escolha for subdividida em duas escolhas sucessivas, a entropia original deverá ser a soma ponderada dos valores individuais de H.

A última propriedade pode ser interpretada do seguinte modo: a entropia é função da distribuição em si e não depende da forma como são agrupados os eventos, isto é, a entropia é uma função de estado. Isto pode ser ilustrado através do exemplo na seguinte figura:

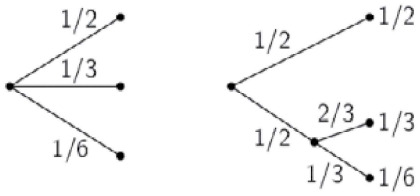

Figura 3. Decomposições de uma escolha com três possibilidades.

Na parte esquerda da figura 3 temos três possibilidades com probabilidades $p_1 = \frac{1}{2},\ p_2 = \frac{1}{3}$ e $p_3 = \frac{1}{6}$. À direita da mesma figura, primeiro escolhemos entre duas possibilidades, cada uma com probabilidade $\frac{1}{2}$, e se a segunda ocorrer, deve-se fazer outra escolha com probabilidades $\frac{1}{3}$ e $\frac{2}{3}$. Os resultados finais terão de ser iguais, isto é, uma função de entropia deverá satisfazer

$$\mathrm{H}\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = \mathrm{H}\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{3}, \frac{2}{3}\right).$$

Shannon provou que uma tal função obedecendo às propriedades 1., 2. e 3. deverá assumir a forma

$$\mathrm{H}(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log_b\ p_i,$$

onde K é uma constante positiva e arbitrária, estando apenas associada à escolha de uma unidade de medida (cf. [5], *Theorem* 2 e *Appendix* 2). A opção mais usual considera $K = 1$ e $b = 2$, subentendendo que a unidade de informação é o bit, obtendo-se nesse caso a expressão (2).

A título de exemplo, $\mathrm{H}(p, q)$, a entropia para o caso de uma variável aleatória com dois estados possíveis, p e q, com $q = 1 - p$, é dada por:

$$\mathrm{H}(p, q) = \mathrm{H}(p, 1 - p) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$

O seu gráfico surge representado na figura seguinte.

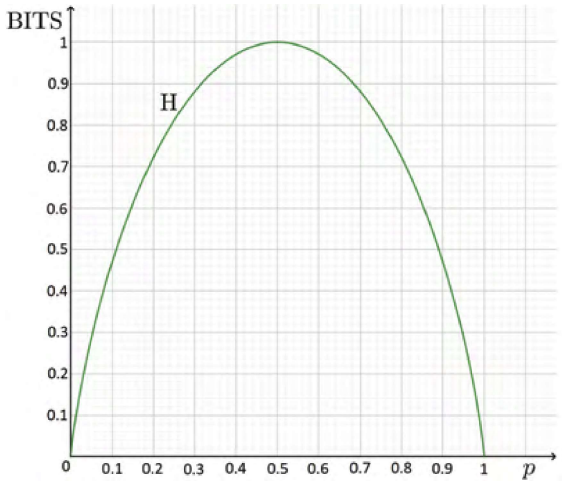

Figura 4. Entropia no caso de dois estados com probabilidades p e $1-p$.

Como expectável, a entropia máxima resulta de $p = q = \frac{1}{2}$, valendo 1 bit.

3. ENTROPIA DE SHANNON E DIVERSIDADE BIOLÓGICA

A quantificação da diversidade ecológica dos ecossistemas é a aplicação biológica mais comum da teoria da informação. Em geral, comunidades biológicas saudáveis localizadas em *habitats* favoráveis tendem a ser vistas como sistemas altamente diversos. Uma diminuição da diversidade biológica poderá dever-se tanto a condições ambientais como a situações de stress na comunidade biológica. A teoria da informação permite quantificar as diferenças entre ecossistemas, tanto as naturais como as

induzidas pelo stress (*e.g.*, devidas a catástrofes ambientais, de origem humana ou não), bem como estudar a evolução da diversidade biológica ao longo do tempo. Os biólogos designam a entropia por índice de Shannon ou diversidade de primeira ordem, não sendo a única ferramenta existente com estes propósitos. Tal índice pode ser calculado para todos os organismos presentes num ambiente ou para tipos específicos de organismos (*e.g.*, árvores ou insetos). Há ainda exemplos de utilização do índice de Shannon no campo da comunicação entre animais. Por exemplo, em [6] o autor apresenta um interessante estudo feito sobre formigas-de-fogo, no qual relaciona a quantidade de informação direcional transmitida a formigas obreiras por um único trilho de odor de formiga-de-fogo, em função da distância entre o ninho e a fonte de alimentação encontrada. Referências a outras aplicações podem ser vistas em [2] .

O exemplo seguinte serve como simples ilustração da utilização desta ferramenta no campo da biologia. Os zoólogos MacArthur e MacArthur [3] propuseram-se estudar as possíveis relações entre a diversidade de espécies de aves (DEA) nidificantes e a diversidade de espécies de plantas (DEP) onde a nidificação ocorre, bem como com certas características dessa vegetação. Fizeram-no em 11 locais de florestas caducifólias de Pennsylvania, Vermont e Maryland. Em particular, além de estimativas para a DEA e para a DEP, uma das características testadas foi a altura das folhagens das árvores, a qual se exprime através do número de camadas de folhas entre o solo e o céu em diferentes locais. Por forma a criar um indicador de diversidade de altura da folhagem (DAF), esta característica foi dividida em três zonas: de zero a dois pés acima do solo, de dois a 25 pés e acima de 25 pés. Os autores obtiveram estimativas para a DAF nos 11 locais e puderam constatar que quando as quantidades de folhas acima do solo nas três zonas são sensivelmente iguais a DAF aumenta, refletindo um ambiente físico mais complexo. Estimados estes três indicadores - DEA, DEP e DAF - nos 11 locais de estudo, os autores procuraram estabelecer conexões entre a atratibilidade de um *habitat* para a nidificação e os referidos indicadores, tendo chegado à conclusão de que existe uma forte correlação entre a DEA e a DAF (ver gráfico na figura 5).

O coeficiente de correlação de Pearson para os dados de DAF e de DEA é:

$$\rho = \frac{COV(DAF, DEA)}{\sqrt{V(DAF)V(DEA)}} \approx 0,95,$$

(onde V e COV designam a variância e a covariância, respetivamente) confirmando uma forte correlação positiva entre DAF e DEA. Os dados aproximam-se bastante da reta de regressão linear:

$$DEA = 2,02 \times DAF + 0,44.$$

Com base nesta relação, as aves parecem selecionar os locais de nidificação com base na maior DAF. Por outro lado, observa-se uma relação muito mais fraca entre os valores de DEA e de DEP. A estrutura física de um bosque em termos das folhagens presentes em diferentes zonas de alturas parece importar mais às aves no momento de nidificar do que propriamente as plantas que produzem tais estruturas.

Local	DAF	DEP	DEA
A	0,043	0,972	0,639
B	0,448	1,911	1,266
C	0,745	2,344	2,265
D	0,943	1,768	2,403
E	0,731	1,372	1,721
F	1,009	2,503	2,739
G	0,577	1,367	1,332
H	0,859	1,776	2,285
I	1,021	2,464	2,277
J	0,825	2,176	2,127
K	1,093	2,816	2,567

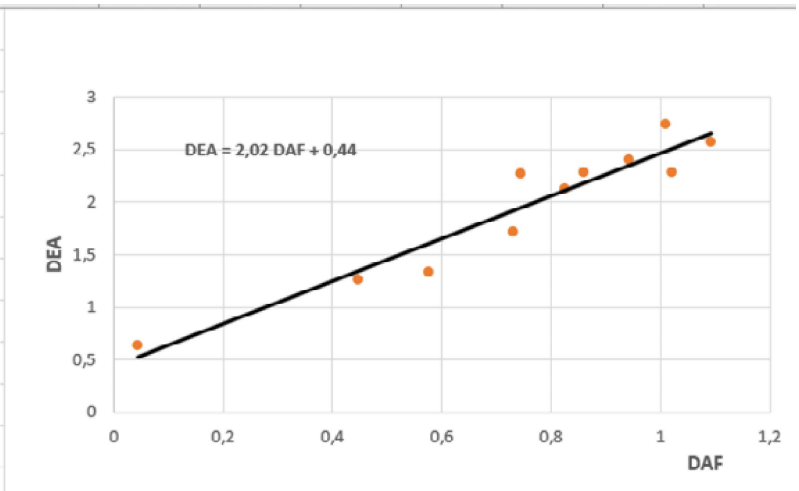

Figura 5. Valores de DAF, DEP e DEA e relação entre DEA e DAF.

4. MIGRAÇÕES DE ESTUDANTES

À semelhança de certas espécies animais, é possível observar fenómenos de migração entre diversos tipos de populações. Em particular, a população estudantil do Ensino Superior migra regularmente desde as respetivas localidades de residência para os estabelecimentos de Ensino Superior (doravante designados por escolas). Fatores como a distância residência-escola, o rendimento familiar e a existência de transportes (públicos ou particulares) influem na opção entre a migração pendular (deslocação diária casa-escola) e a migração sazonal (a qual implica o arrendamento/compra de uma residência no local de estudo). Independentemente desta opção, pode ser relevante determinar quão diversificado é o conjunto de origens geográficas dos alunos de uma escola, como forma de, por exemplo, ajustar medidas pedagógicas ou políticas de fomento da atratibilidade dos cursos. Ao incluir o local de residência, o registo de matrícula de cada aluno permite atingir tal propósito com bastante fiabilidade.

Comecemos por agrupar os alunos de um dado universo (turma, disciplina, curso ou escola) em categorias de acordo com o respetivo local de residência. Para tal, consideramos uma categoria por cada região NUTS III. NUTS é o acrónimo de Nomenclatura das Unidades Territoriais para Fins Estatísticos, sistema hierárquico de divisão do território em regiões. A nomenclatura subdivide-se em três níveis (NUTS I, NUTS II e NUTS III), definidos de acordo com critérios populacionais, administrativos e geográficos. Assim, os 308 atuais municípios de Portugal agrupam-se em 25 NUTS III (subdivisões de sete NUTS II e de três NUTS I). Além disso, e porque é frequente haver estudantes estrangeiros, consideramos três categorias adicionais: Brasil, PT e Europa, para estudantes provenientes do Brasil, dos PALOPs/Timor-Leste e da Europa, respetivamente. No total, os alunos foram agrupados em 28 categorias. Evidentemente, categorias adicionais podem sempre ser consideradas (agrupando, por exemplo, alunos de outros continentes ou subdividindo as duas últimas categorias). A medida de diversidade aplicada no presente contexto, constituindo uma métrica para a diversidade de origens geográficas dos estudantes, resultará então do cálculo de

$$\mathrm{H}\,(p_1,\ldots,p_{28}) = -\sum_{i=1}^{28} p_i \log_2\ p_i,$$

onde p_i representa a proporção de estudantes pertencentes à i-ésima categoria (a ordem é irrelevante). De acordo com a convenção (3), para cada região j sem alunos, considerou-se $p_j \log_2 p_j = 0$. A medida foi aplicada ao universo de 280 inscritos em 2021/2022 na disciplina de Cálculo I da Licenciatura em Economia da U. de Coimbra. Os dados obtidos e o valor de H (na célula a verde) surgem na tabela que se segue.

NUTS III	NUTS II	NUTS I	Alunos	Proporção (p_i)	$p_i \log_2(p_i)$
Alto Minho	Norte	Continente	11	0,04	-0,18
Cávado			10	0,04	-0,17
Ave			9	0,03	-0,16
Área Metropolitana do Porto			22	0,08	-0,29
Alto Tâmega			3	0,01	-0,07
Tâmega e Sousa			20	0,07	-0,27
Douro			6	0,02	-0,12
Terras de Trás-os-Montes			5	0,02	-0,10
Oeste	Centro		6	0,02	-0,12
Região de Aveiro			13	0,05	-0,21
Região de Coimbra			66	0,24	-0,49
Região de Leiria			13	0,05	-0,21
Viseu Dão Lafões			13	0,05	-0,21
Beira Baixa			1	0,00	0,00
Médio Tejo			5	0,02	-0,10
Beiras e Serra da Estrela			6	0,02	-0,12
Área Metropolitana de Lisboa	Área Metropolitana de Lisboa		10	0,04	-0,17
Alentejo Litoral	Alentejo		0	0,00	0,00
Baixo Alentejo			0	0,00	0,00
Lezíria do Tejo			7	0,03	-0,13
Alto Alentejo			1	0,00	-0,03
Alentejo Central			0	0,00	0,00
Algarve	Algarve		3	0,01	-0,07
Região Autónoma dos Açores	Região Autónoma dos Açores	Região Autónoma dos Açores	1	0,00	0,00
Região Autónoma da Madeira	Região Autónoma da Madeira	Região Autónoma da Madeira	3	0,01	-0,07
Outras Categorias					
PT (PALOPs+Timor)			30	0,11	-0,35
Brasil			16	0,06	-0,24
Europa			0	0,00	0,00
			280	1,00	3,87

Figura 6. Diversidade de origem geográfica de população estudantil.

A análise da tabela na figura 6 permite observar uma expectável predominância de alunos provenientes da região de Coimbra, acompanhada de outras regiões com alguma relevância relativa. Se em algumas tal é compreensível (caso das regiões fronteiras à de Coimbra), noutras contudo ocorre alguma surpresa (*e.g.*, algumas regiões inseridas na NUTS II do Norte contribuem com proporções assinaláveis). Destaca-se igualmente o peso dos alunos das categorias PT e Brasil. Observe-se que a diversidade máxima seria de $\mathrm{H}_{max} = \log_2\ 28 \approx 4,81$, pelo que a diversidade de origem geográfica se cifrou em cerca de 81% do seu máximo. A recolha anual destes dados permite acompanhar este indicador ao longo dos anos. Assim, por exemplo, considerando os anos letivos 2009/2010 (o primeiro para o qual a plataforma informática fornece os dados necessários), 2020/2021 e 2021/2022, e para as mesmas 28 categorias (com, respetivamente, 420, 280 e 280 alunos),

observa-se um claro aumento da diversidade de origens geográficas dos alunos no estrito universo considerado.

Ano Lectivo	2009/2010	2020/2021	2021/2022
H	2,86	3,66	3,87

São diversas as possibilidades de aplicação deste indicador, desde estudos comparativos entre cursos ou entre faculdades (ou até mesmo entre universidades, confirmando ou refutando tendências de regionalização de uma escola), passando pelo estabelecimento de séries temporais (por curso, por exemplo). Evidentemente, o que tornará mais interessante a sua utilização será, como acima se referiu, o modo como ele permitirá eventualmente ajustar medidas pedagógicas ou políticas de atratibilidade.

O autor agradece à Cristina Martins (DMUC) a leitura cuidadosa do texto e as sugestões de redação de parte da secção 2.

REFERÊNCIAS

[1] Hartley, R. V. L., "Transmission of information."}*The Bell System Technical Journal*, vol. 7, no. 3, pp. 535-563, July 1928. doi: 10.1002/j.1538-7305.1928.tb01236.x.

[2] Kolmes, S. and Mitchell, K., "Information Theory and Biological Systems." *UMAP Modules: Tools for Teaching* 1990, pp. 43-78, UMAP Module 705.

[3] MacArthur, Robert H. and John W. MacArthur, "On Bird Species Diversity." *Ecology*, vol. 42, no. 3, pp. 594–98, 1961 JSTOR, *https://doi.org/10.2307/1932254*.

[4] Nyquist, H., "Certain Factors Affecting Telegraph Speed." *Transactions of the American Institute of Electrical Engineers*, vol. XLIII, pp. 412-422, January-December 1924. doi: 10.1109/T-AIEE.1924.5060996.

[5] Shannon, C. E., "A Mathematical Theory of Communication." *The Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, July, October, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x

[6] Wilson, E. O., "Chemical Communication Among Workers of the Fire Ant Solenopsis Saevissima. (Fr. Smith) 2. An Information Analysis of the Odor Trail." *Animal Behaviour*, vol. 10, 1962, pp. 134–147.

[7] *https://www.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/information-entrop*

SOBRE O AUTOR

Paulo Saraiva é professor na Faculdade de Economia da Universidade de Coimbra (FEUC). Licenciado e mestre em Matemática (especialização em Ensino) pela FCTUC, doutorou-se em Economia Matemática e Modelos Econométricos pela FEUC em 2004. É membro da Linha de Álgebra e Combinatória do *CMUC - Centre for Mathematics of the University of Coimbra* e colaborador do *CeBER - Centre for Business and Economics Research*. É coeditor do Boletim *FEUC et al.* desde o início desta publicação.