UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Higor Yudi Duenha Sigaki

Física Estatística e Aprendizagem de Máquina Aplicadas ao Estudo de Sistemas Complexos

Maringá, Fevereiro de 2022.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Higor Yudi Duenha Sigaki

Física Estatística e Aprendizagem de Máquina Aplicadas ao Estudo de Sistemas Complexos

Tese de doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, Fevereiro de 2022.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

S574f

Sigaki, Higor Yudi Duenha
Física estatística e aprendizagem de máquina aplicadas ao estudo de sistemas complexos / Higor Yudi Duenha Sigaki. -- Maringá, PR, 2022.
123 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro.
Tese (Doutorado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-Graduação em Física, 2022.

1. Análise de dados. 2. Sistemas complexos. 3. Física estatística. 4. Aprendizagem de máquina. 5. Ciência de dados. I. Ribeiro, Haroldo Valentin, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física. III. Título.

CDD 23.ed. 530.13

Elaine Cristina Soares Lira - CRB-9/1202

HIGOR YUDI DUENHA SIGAKI

FÍSICA ESTATÍSTICA E APRENDIZAGEM DE MÁQUINA APLICADAS AO ESTUDO DE SISTEMAS COMPLEXOS

Tese apresentada à Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de doutor.

Aprovado em: Maringá, 17 de fevereiro de 2022.

BANCA EXAMINADORA

_____

Prof. Dr. Haroldo Valentin Ribeiro
Universidade Estadual de Maringá – UEM

_____

Prof. Dr. Osvaldo Anibal Rosso
Universidade Federal de Alagoas – UFAL

_____

Prof. Dr. José Soares de Andrade Júnior
Universidade Federal do Ceará – UFC

_____

Prof. Dr. Rafael Soares Zola
Universidade Tecnológica Federal do Paraná – UTFPR/Apucarana
Programa de Pós-Graduação em Física - PFI/UEM

_____

Prof. Dr. Renio dos Santos Mendes
Universidade Estadual de Maringá – UEM

Agradecimentos

Muitas vezes só nos damos conta do que passamos em alguma situação, ainda mais quando essa situação dura por um longo período de tempo, ao olhar pra trás e perceber tudo o que contribuiu para que chegássemos até aqui. Ao escrever essa mensagem de agradecimentos como a última seção a ser adicionada nessa tese, esse é o sentimento que me toma nesse momento. Um sentimento, claro, de dever cumprido, mas sobretudo de satisfação a todo aprendizado e crescimento que tive o privilégio de vivenciar durante os últimos cinco anos. Espero que esse crescimento tenha sido, principalmente, como ser humano.

Nem tudo são flores e na academia não seria diferente, mas tive a sorte de ter como referência um orientador que desempenha seu papel com maestria. Uma pessoa que, para mim, é uma das maiores referências em pesquisa científica. Muito obrigado, Haroldo! Pela oportunidade de trabalhar com você e, principalmente, por todo apoio em absolutamente todas as situações. Sempre tive claro comigo que poderia pedir sua ajuda ou um conselho e de que teria uma resposta pertinente e considerada.

Agradeço também a minha família, a minha mãe Maria, ao meu pai Hélio e aos meus irmãos, Débora, Eduardo e Hugo por sempre me incentivarem, cada um a sua maneira. A minha noiva Larissa, por ter estado ao meu lado desde sempre (desde o segundo ano da graduação, pra ser mais preciso) e ter me apoiado e me ajudado a chegar até aqui.

Não poderia deixar de agradecer imensamente aos meus amigos do ComplexLab: Alvaro, André, Arthur, Denner, Diego, Fera, Gustavo, Leandro, Max e Prof. Renio, que são de extrema importância, principalmente por proporcionarem momentos valiosos de descontração e muitas conversas interessantes. Aprendi muito com cada um de vocês!

Também quero agradecer aos colaboradores com quem fui coautor dos artigos apresentados nessa tese: Luiz Gustavo, Renato, Prof. Rafael Zola, Prof. Rodolfo de Souza, Prof. Ervin Lenzi e Prof. Matjaž Perc. Pesquisadores cujas competência, criatividade e disposição são indiscutíveis.

E por último, mas tão importante quanto todos, agradeço a Deus por minha vida, família e amigos.

Great things are not done by impulse,
but by a series of small things brought together.
Vincent van Gogh

Resumo

Neste trabalho, investigamos diferentes sistemas complexos empregando métodos de Física Estatística combinados com técnicas de Ciência de Dados. Na maioria dos estudos apresentados, utilizamos duas medidas de complexidade: a entropia e a complexidade estatística de permutação. Técnicas de aprendizado de máquina são combinadas a essas quantidades para extrair propriedades emergentes, tendências e padrões que resultam das interações entre os elementos individuais do sistema. Em particular, utilizamos imagens de obras de arte de uma grande base de dados para quantificar conceitos qualitativos propostos por historiadores da arte e discriminar entre períodos da arte e estilos diferentes. Em outro estudo, propomos duas abordagens para extrair propriedades físicas de cristais líquidos diretamente a partir das imagens das texturas desses materiais. Finalmente, em um último estudo, investigamos a hipótese de mercado eficiente no mercado de criptomoedas e no mercado de ações. Nossos resultados demonstram que abordagens simples inspiradas na Física podem ser combinadas com ferramentas da Ciência de Dados para investigar uma grande variedade de disciplinas, incluindo Física experimental, História da Arte e Economia.

Palavras-chave: Sistemas complexos. Ciência de Dados. Aprendizagem de máquina. Entropia. Complexidade. História da Arte. Cristais líquidos. Criptomoedas. Mercado de ações.

Abstract

In this work, we investigate different complex systems by employing concepts from Statistical Physics combined with Data Science techniques. In the majority of the studies presented here, we rely on two complexity measures, namely, permutation entropy and statistical complexity. Machine learning techniques are employed in combination with these quantitites to extract emergent properties, trends and patterns, which results from the interactions among the individual elements of the system. In particular, by using digital images of artworks from a large database, we quantify qualitative concepts proposed by art historians and discriminate among different art periods and styles. In another study, we propose two approaches to extract physical properties of liquid crystals directly from texture images of these materials. Finally, in a last study, we probe the efficient market hypothesis on the cryptocurrency market and on the stock market. Our results thus demonstrate that physics-inspired approaches can be successfully combined with Data Science tools for investigating a wide range of disciplines, including experimental physics, art history and economics.

Keywords: Complex systems. Data Science. Machine learning. Entropy. Complexity. History of Art. Liquid crystals. Cryptocurrencies. Stock market.

Sumário

Introdução

O termo *"big data"* tem se destacado cada vez mais desde o seu surgimento nos anos 1990 [1]. Da maneira como foi concebida, essa expressão era utilizada para denominar conjuntos de dados extremamente grandes, que surgiam devido ao rápido e amplo desenvolvimento tecnológico observado ao longo desse mesmo período. Atualmente, é fato que dispomos de um gigantesco volume de dados, das mais variadas naturezas, sobre os mais diversos sistemas e, principalmente, em um grau de detalhe impressionante e até inimaginável poucas décadas atrás. Por dados, entendemos qualquer tipo de informação mensurável ou descritiva sobre um objeto, sistema ou situação. Esses dados têm as mais diversas origens, incluindo informações registradas por sensores, transações de compras, aparelhos celulares, publicações nas redes sociais, imagens digitais e até interações sociais, sejam elas de natureza humana ou animal. Para fornecer uma noção do volume de dados que produzimos e da taxa impressionante com que esse volume cresce, evidenciamos dois números que se destacam. Estimativas revelam que, todos os dias, são criados 2,5 quintilhões de *bytes* de dados no mundo [2] e, além disso, 90% de todos os dados disponíveis atualmente foram produzidos somente ao longo dos últimos dois anos [3].

Muitos afirmam que as revoluções na ciência tendem a ser precedidas por revoluções na medição [4]. Da mesma maneira que a invenção do microscópio viabilizou importantes avanços científicos em diversas direções, a capacidade de produzir e analisar dados na era do *big data*, permitiu à ciência estudar vários sistemas em um nível mais geral. Um exemplo é o estudo do comportamento humano, o qual hoje pode ser realizado com grande grau de detalhe em toda uma população. Junto com essa revolução do *big data*, veio a necessidade de profissionais com as habilidades e a *expertise* para transformar esses dados em informações de valor: o chamado *cientista de dados*.

De acordo com o LinkedIn, essa carreira aparece no *ranking* das posições mais emergentes nos Estados Unidos por três anos consecutivos (2018-2020), com uma taxa de crescimento

anual média de 35% [5]. No Brasil, além de também figurar no *ranking* das posições mais emergentes, essa taxa de crescimento anual da oferta de vagas para cientistas de dados chega a 78% [6]. Esse profissional tem como objetivo organizar e analisar esses grandes conjuntos de dados, produzindo *insights* importantes sobre as mais diversas questões. Muitos desses problemas, por sua vez, são extremamente críticos e balizam decisões de grande impacto como a criação e o aprimoramento de políticas governamentais, as quais, geralmente, afetam grande parte da população.

A capacidade de analisar sistemas que, por vezes, jamais haviam sido estudados em larga escala, fez surgir novas áreas de pesquisa e da ciência. Dentre os diversos exemplos podemos citar: econofísica [7], estudo do comportamento coletivo de animais [8] e das mudanças climáticas globais [9], bem como a sociofísica [10]. Um fato que corrobora a expansão desse campo fértil de pesquisa envolvendo grandes bases de dados é a publicação frequente desses estudos nas mais renomadas revistas científicas. Estudos sobre o espalhamento de vírus e epidemias [11–13], monitoramento do meio ambiente [14], quantificação da reputação nas artes visuais [15], mapeamento do funcionamento do governo em nível estadual [16] e a identificação de tipos de personalidade humana [17] são apenas alguns exemplos recentes da diversidade de questões que estão sendo abordadas.

Esses estudos têm em comum a busca por padrões, tendências e associações que resultam das interações e relações entre as componentes individuais do conjunto de dados. Revelar essas propriedades *emergentes* é um dos principais propósitos da Física de Sistemas Complexos e o foco dos estudos desenvolvidos em nosso grupo de pesquisa (`complex.pfi.uem.br`) [18–27]. De fato, o *framework* teórico da Física de Sistemas Complexos, com suas abordagens e metodologias, é bastante adequado para modelar como os elementos de um conjunto de dados interagem e influenciam uns aos outros. Portanto, percebemos que os objetivos da Ciência de Dados e do estudo de sistemas complexos são amplamente convergentes. Embora os conjuntos de dados relacionados à investigações em sistemas complexos muitas vezes não sejam estritamente classificados como *"big data"*, uma abordagem tradicional certamente não seria viável ou nem mesmo possível na maioria dos casos.

Nesse sentido, essa tese apresenta alguns estudos sobre sistemas complexos que empregam conceitos de Física Estatística combinados com técnicas de Ciência de Dados. Iniciamos nossa apresentação no capítulo 1 descrevendo os métodos que fundamentam todas as nossas análises. Como o leitor irá notar, esse trabalho tem por essência o uso de duas medidas de complexidade: a entropia de permutação [28] e a complexidade estatística de permutação [29, 30]. Juntamente com essas medidas, combinamos técnicas de aprendizagem estatística [31] para investigar comportamentos coletivos e emergentes de quatro sistemas complexos. No capítulo 2, apresentamos os resultados de um estudo em que utilizamos imagens de obras de arte para quantificar conceitos qualitativos da história da arte [32]. No capítulo 3, propomos duas abordagens para extrair propriedades físicas de cristais líquidos

a partir das imagens de suas texturas [33, 34]. A primeira, utiliza as medidas de complexidade combinadas com métodos de aprendizagem estatística. Já a segunda abordagem, emprega métodos de aprendizagem de máquina – redes convolucionais neurais – diretamente nas imagens das texturas desses materiais. No capítulo 4, investigamos aspectos da teoria do mercado eficiente ao quantificar a eficiência dos mercados de criptomoedas [35] e de ações [36] por meio de suas séries financeiras. Por fim, apresentamos nossas conclusões e perspectivas para trabalhos futuros.

CAPÍTULO 1

Métodos estatísticos para análise de dados

Neste capítulo, apresentamos os métodos que foram utilizados ao longo do desenvolvimento desse trabalho visando identificar padrões em sistemas complexos. Iniciamos essa apresentação descrevendo duas medidas de complexidade estatística conhecidas como entropia de permutação e complexidade estatística de permutação [28–30, 37]. Em seguida, discutimos alguns conceitos relacionados aos métodos de aprendizado de máquina (*machine learning*).

1.1 A entropia de permutação e a abordagem de Bandt e Pompe

A entropia de permutação é um método proposto por Bandt e Pompe [28] que visa, fundamentalmente, encontrar uma medida de complexidade "natural" para séries temporais arbitrárias. O problema, segundo argumentam, é que existem várias medidas e significados diferentes para o conceito de complexidade que permitem comparar séries temporais e distingui-las entre diversas características, por exemplo, entre regular e caótica. Entretanto, a maioria dessas medidas não pode ser aplicada para uma série temporal arbitrária, ou seja, cada tipo de série demanda um algoritmo específico para extrair a medida de complexidade em questão. Essa, por sua vez, usualmente depende de parâmetros específicos e ajustáveis, impossibilitando que resultados sejam obtidos sem o conhecimento de pormenores dos métodos e dificultando a reprodutibilidade científica ainda mais [38–41].

A fim de encontrar uma medida de complexidade que possa ser aplicada a qualquer série temporal e que seja também robusta, simples e computacionalmente rápida, Bandt

e Pompe [28] propuseram uma medida de complexidade estatística para séries temporais chamada *entropia de permutação*.

Antes de definirmos formalmente essa medida, vamos ilustrá-la por meio de um exemplo. Para isso, consideramos uma série com $n = 7$ termos definida por

$$x = \{4, 7, 9, 10, 6, 11, 3\}.$$

Vamos, inicialmente, particionar essa série em parcelas superpostas de tamanho $d = 2$, sendo d o tamanho da partição que é usualmente conhecido como *embedding dimension*. Podemos representar essas $(n - d + 1) = 6$ partições pelos seguintes vetores:

$$\begin{aligned}
(\vec{2}) &= (4, 7), \\
(\vec{3}) &= (7, 9), \\
(\vec{4}) &= (9, 10), \\
(\vec{5}) &= (10, 6), \\
(\vec{6}) &= (6, 11), \\
(\vec{7}) &= (11, 3).
\end{aligned}$$

Em seguida, vamos analisar a ordem entre os elementos desses vetores, obtendo os dois grupos listados no quadro abaixo:

permutação “01”	permutação “10”
$(\vec{2}) = (4, 7)$	$(\vec{5}) = (10, 6)$
$(\vec{3}) = (7, 9)$	$(\vec{7}) = (11, 3)$
$(\vec{4}) = (9, 10)$	
$(\vec{6}) = (6, 11)$	

No primeiro grupo, temos os vetores cuja primeira componente é menor que a segunda (permutação “01”). Já no outro grupo, a primeira componente é maior que a segunda (permutação “10”). Para representar essas duas condições, utilizamos os símbolos “01” quando a ordem é crescente e “10” quando ela é decrescente. Assim, nesse exemplo, temos quatro pares do tipo “01” e dois do tipo “10”. Logo, para essa série, podemos definir a probabilidade de encontrar vetores do tipo “01” como $p(\text{“01”}) = 4/6$ e do tipo “10” como $p(\text{“10”}) = 2/6$.

A entropia de permutação para essa série é, simplesmente, a entropia de Shannon [42] calculada para essas probabilidades. Notamos que o termo “permutação” diz respeito ao método de simbolização proposto por Bandt e Pompe [28] para obter a distribuição de probabilidade relacionada aos estados acessíveis do sistema em questão. Lembrando que a

entropia de Shannon associada a uma distribuição de probabilidades $\{p_1, p_2, \ldots, p_m\}$ é[1]

$$S = -\sum_{i=1}^{m} p_i \log p_i, \tag{1.1}$$

escrevemos a entropia de permutação como

$$S = -\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right) \approx 0{,}918. \tag{1.2}$$

Do mesmo modo que a entropia de Shannon é uma medida do "grau de desordem" de um sistema, a entropia de permutação mede a "desordem" na ocorrência dos padrões "01" e "10" ao longo da série que estamos investigando, ou seja, ela quantifica a dinâmica de ordenamento de pares consecutivos de elementos da série.

Uma outra possibilidade é tomar partições de tamanho $d = 3$. Nesse caso, obtemos os seguintes $(n - d + 1) = 5$ vetores:

$$\begin{aligned}
(\vec{3}) &= (4, 7, 9),\\
(\vec{4}) &= (7, 9, 10),\\
(\vec{5}) &= (9, 10, 6),\\
(\vec{6}) &= (10, 6, 11),\\
(\vec{7}) &= (6, 11, 3),
\end{aligned}$$

os quais podem ser agrupados em $3! = 6$ categorias associadas ao ordenamento de seus elementos. Em particular, os vetores $(\vec{3})$ e $(\vec{4})$ são representados pela permutação "012", já que seus elementos estão na ordem crescente. Os vetores $(\vec{5})$ e $(\vec{7})$ correspondem à permutação "201", pois $a_2 < a_0 < a_1$, se for utilizada a notação (a_0, a_1, a_2) para representar cada partição. Já o vetor $(\vec{6})$ é representado pela permutação "102" porque $a_1 < a_0 < a_2$. Observamos que para $d = 3$, das 6 permutações possíveis, apenas três estão presentes no exemplo em questão, conduzindo às seguintes probabilidades:

$$\begin{aligned}
p(\text{“012”}) &= 0{,}4,\\
p(\text{“021”}) &= 0,\\
p(\text{“102”}) &= 0{,}2,\\
p(\text{“120”}) &= 0,\\
p(\text{“201”}) &= 0{,}4,\\
p(\text{“210”}) &= 0,
\end{aligned}$$

em que $p(\pi_i)$ representa a probabilidade de ocorrência da permutação π_i. Dessa forma, a

[1] Ao longo desse texto, vamos considerar o logaritmo na base e para o cálculo da entropia e a constante multiplicativa como sendo unitária, isto é, $k = 1$ em $S = -k\sum_{i=1}^{m} p_i \log p_i$. Essa escolha não acarreta em perda de generalidade, pois usaremos a versão normalizada da entropia, como na equação 1.11.

entropia de permutação para $d = 3$ é

$$S = -2\left(\frac{2}{5}\right)\log\left(\frac{2}{5}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right) \approx 1{,}522. \tag{1.3}$$

Ilustrada a técnica proposta por Bandt e Pompe via um exemplo, vamos agora apresentá-la formalmente. Para isso, consideramos uma série temporal genérica composta de n elementos e representada por

$$\{x_1, x_2, \ldots, x_n\} = \{x_t\}_{t=1,2,\ldots,n}. \tag{1.4}$$

Tomamos partições de tamanho d nessa série, as quais são representadas pelos vetores

$$(\vec{s}) \mapsto (x_{s-(d-1)}, x_{s-(d-2)}, \ldots, x_{s-1}, x_s), \tag{1.5}$$

com $s = d, d+1, \ldots, n$. Para cada um desses $(n-d+1)$ vetores, estudamos todas as $d!$ permutações π_i de ordem d, que são entendidas como as possíveis maneiras de ordenamento entre d símbolos, isto é, as possíveis permutações do vetor $(0, 1, \ldots, d-1)$. Em seguida, determinamos a frequência relativa (probabilidade de ocorrência) de cada uma dessas $d!$ permutações, representada por

$$p(\pi_i) = \frac{\#\{s | s \leq (n-d+1);\ (\vec{s}) \text{ do tipo } \pi_i\}}{n-d+1}, \tag{1.6}$$

na qual o símbolo # refere-se a cardinalidade do conjunto, ou seja, o número de ocorrências da permutação π_i. Com isso, temos o conjunto das probabilidades $P = \{p(\pi_i)\}_{i=1,2,\ldots,d!}$ que é utilizado para determinar a entropia de permutação de ordem d associada à série temporal $\{x_t\}_{t=1,2,\ldots,n}$, isto é,

$$S[P] = -\sum_{i=1}^{d!} p(\pi_i)\log p(\pi_i). \tag{1.7}$$

Vemos que o único parâmetro usado para determinar a entropia de permutação é a *embedding dimension* d. Esse parâmetro tem grande importância, já que $d!$ representa o número de estados acessíveis ao sistema. Em seu trabalho, Bandt e Pompe recomendam que $d = (3, \ldots, 7)$ para a maior parte das aplicações com séries empíricas, pois para que a estimativa de P seja estatisticamente confiável, a relação $d! \ll n$ deve ser satisfeita.

Em posse da definição dada pela equação 1.7, podemos calcular a entropia de permutação normalizada

$$H[P] = \frac{S[P]}{S_{max}}, \tag{1.8}$$

na qual S_{max} representa o valor máximo da entropia de Shannon que é obtido ao considerarmos que todos os estados acessíveis ao sistema são equiprováveis. Qualitativamente isso

significa que nosso conhecimento sobre qual estado o sistema será encontrado é mínimo, portanto, a entropia, que nesse contexto é interpretada como a ignorância, é máxima. De fato, podemos mostrar que a distribuição de probabilidade "menos informativa" e que maximiza a entropia é a distribuição uniforme

$$P_e = \{p(\pi_i) = 1/d!\,,\ i = 1, \ldots, d!\}, \tag{1.9}$$

conduzindo ao valor

$$S_{max} = S[P_e] = \log d!. \tag{1.10}$$

Usando esse resultado, a entropia de permutação normalizada é escrita como

$$H[P] = \frac{S[P]}{S_{max}} = \frac{S[P]}{\log d!}. \tag{1.11}$$

Observamos que após a normalização, os valores da entropia H ficam limitados ao intervalo $0 \leq H[P] \leq 1$, no qual o limite inferior, $H[P] = 0$, é obtido quando a série é completamente regular. Já o caso $H[P] = 1$ ocorre para uma série completamente aleatória, na qual todas as $d!$ permutações possíveis são equiprováveis.

1.2 Outras medidas de complexidade estatística

A entropia de permutação proposta por Bandt e Pompe é uma medida de complexidade muito útil para capturar informações a respeito da dinâmica de sistemas complexos. Porém, ela não é suficiente para capturar uma noção mais intuitiva de complexidade. Como exemplo dessa limitação, consideramos dois sistemas físicos típicos: o cristal perfeito e o gás ideal. Tais sistemas são bem descritos por modelos teóricos e apresentam leis empíricas bem estabelecidas. Do ponto de vista macroscópico, são amplamente previsíveis e, portanto, são sistemas de baixa complexidade. Entretanto, do ponto de vista da entropia eles representam situações extremas e completamente opostas. O cristal é altamente regular/ordenado e tem entropia baixa. Já o gás ideal é completamente desordenado e caracterizado por uma entropia alta.

Esperamos encontrar uma medida de complexidade estatística que satisfaça as noções intuitivas apontadas anteriormente. Essa medida pode ser obtida levando em conta o que López-Ruiz *et al.* [29] definem como "desequilíbrio" $\mathcal{D}$, o qual é fundamentalmente uma medida da distância entre a distribuição equiprovável e a distribuição de probabilidade dos estados acessíveis ao sistema. Qualitativamente, esse "desequilíbrio" faz alusão à existência de uma hierarquia entre as probabilidades do sistema. No caso de existirem estados privilegiados ou mais prováveis dentre os acessíveis, $\mathcal{D}$ será diferente de zero.

A título de ilustração, no caso do cristal perfeito, no qual os átomos estão completa-

mente ordenados, a distribuição de probabilidade dos estados acessíveis está centrada em torno de um único estado devido à simetria. Nesse caso, observamos uma hierarquia clara e, ao calcularmos o "desequilíbrio," obtemos um valor máximo. Já no caso de um gás ideal, o sistema pode ser encontrado em qualquer um dos estados acessíveis com a mesma probabilidade. Assim, o desequilíbrio é mínimo, indicando a ausência de uma hierarquia entre as probabilidades, ou seja, todas têm o mesmo peso com relação ao possível estado em que o sistema pode ser encontrado.

Dessa forma, observando a figura 1.1, percebemos que, sozinhas, as medidas de entropia e desequilíbrio não são capazes de quantificar a ideia intuitiva de complexidade estatística para o cristal e para o gás ideal. A solução para esse problema, proposta por López-Ruiz *et al.* [29], foi considerar o produto entre a entropia H e o desequilíbrio $\mathcal{D}$, uma quantidade que possui as características requeridas anteriormente, ou seja, tende a zero tanto para o cristal quanto para o gás ideal e é diferente de zero para outros sistemas.

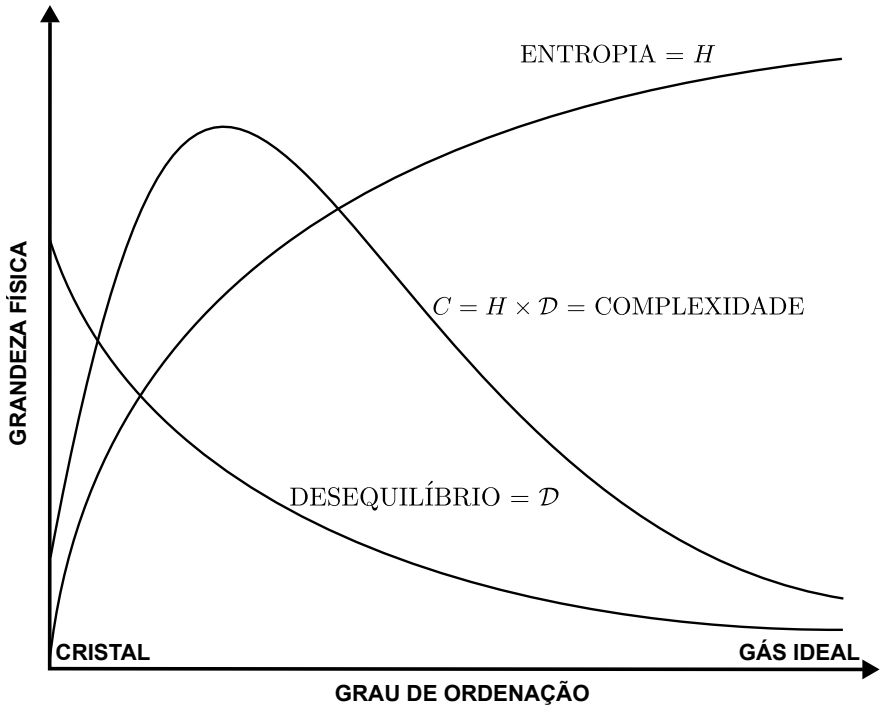

Figura 1.1: Ilustração do comportamento intuitivo esperado para as grandezas entropia H, desequilíbrio $\mathcal{D}$ e complexidade $C = H\mathcal{D}$, todas em função do grau de ordenação do sistema. No extremo de ordem podemos ter um cristal; já no extremo de aleatoriedade, um gás ideal. Notamos que a medida de complexidade estatística de López-Ruiz *et al.* [29] captura a noção intuitiva de complexidade para os exemplos citados, ou seja, tende a zero para ambos. Figura adaptada da referência [29].

Desse modo, a partir do trabalho desenvolvido por López-Ruiz *et al.* [29], obtemos uma medida de complexidade estatística que, agregada à prescrição de Bandt e Pompe, pode trazer à tona informações sobre a dinâmica dos sistemas que não são capturadas apenas por medidas de aleatoriedade/ordem, como a entropia de permutação. Matematicamente, a

complexidade proposta por López-Ruiz *et al.* é definida como

$$C[P] = H[P]\mathcal{D}[P] = -C_0 \left(\sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i) \right) \left(\sum_{i=1}^{d!} \left(p(\pi_i) - \frac{1}{d!} \right)^2 \right), \tag{1.12}$$

com $H[P]$ sendo a entropia, C_0 uma constante de normalização e $P = \{p(\pi_i)\}_{i=1,2,\ldots,d}$ o conjunto das probabilidades dos estados acessíveis i ao sistema. Observamos que $\mathcal{D}[P]$ é simplesmente a distância euclidiana entre a distribuição de probabilidade do sistema e a distribuição equiprovável ($P_e = \{p(\pi_i) = 1/d!\,,\, i = 1, \ldots, d!\}$).

Relembrando dos casos extremos ilustrados anteriormente, vemos que essa medida de complexidade é plausível. Para um cristal, o desequilíbrio $\mathcal{D}$ é grande por causa da presença de um estado privilegiado, mas a entropia H é pequena em razão da simetria, então $C \to 0$. Por outro lado, H é grande e o desequilíbrio $\mathcal{D}$ é mínimo para um gás ideal, já que os estados são equiprováveis, então $C \to 0$ da mesma forma.

Na verdade, essa interpretação intuitiva de complexidade também pode ser obtida por meio de outras medidas entrópicas e de desequilíbrio existentes na literatura [43, 44]. Além das já mencionadas entropia de Shannon [42]

$$H[P] = -\sum_{i}^{d!} p(\pi_i) \log p(\pi_i), \tag{1.13}$$

e da distância euclidiana

$$\mathcal{D}[P] = \sum_{i}^{d!} \left(p(\pi_i) - \frac{1}{d!} \right)^2, \tag{1.14}$$

podemos listar:

- Entropia de Tsallis [45, 46]:

$$H_q[P] = \frac{1}{q-1} \left[1 - \sum_{i}^{d!} (p(\pi_i))^q \right], \tag{1.15}$$

  com q sendo um parâmetro real.

- Entropia de Rényi [47, 48]:

$$H_\alpha[P] = \frac{1}{1-\alpha} \log \left[\sum_{i}^{d!} (p(\pi_i))^\alpha \right], \quad (\alpha \geq 0 \text{ e } \alpha \neq 1), \tag{1.16}$$

  com α sendo um parâmetro real.

- Distância de Wooters [49]:

$$\mathcal{D}[P] = \cos^{-1}\left\{\sum_{i}^{d!}(p(\pi_i))^{1/2}\left(\frac{1}{d!}\right)^{1/2}\right\}. \tag{1.17}$$

- Divergência de Jensen-Shannon [50]:

$$\mathcal{D}[P] = \mathcal{D}_0\{H[(P+P_e)/2] - H[P]/2 - H[P_e]/2\}, \tag{1.18}$$

  com H sendo a entropia de Shannon (equação 1.13) e $\mathcal{D}_0$ uma constante de normalização.

Uma característica importante da medida de complexidade proposta por López-Ruiz *et al.* [29] (equação 1.12) é o fato dela não ser uma função unívoca da entropia [43, 51]. Isso significa que para um dado valor de entropia H existe um intervalo de valores possíveis para a complexidade C, como podemos observar na figura 1.2. Nessa figura, usamos um exemplo para ilustrar esse fato. Por simplicidade, consideramos um sistema com apenas 3 estados acessíveis. Além disso, vamos assumir que a distribuição de probabilidade desses estados seja representada por $P = \{a, b, 1-(a+b)\}$ com $a > 0$, $b > 0$ e $(a+b) \leq 1$. Nesse caso, a entropia é dada por

$$H = -(a\log a + b\log b + [1-(a+b)]\log[1-(a+b)]),$$

já o desequilíbrio $\mathcal{D}$ é

$$\mathcal{D} = \left(a - \frac{1}{3}\right)^2 + \left(b - \frac{1}{3}\right)^2 + \left([1-(a+b)] - \frac{1}{3}\right)^2.$$

Para evidenciar que a correspondência entre H e C não é um a um, notamos que tomando

$$P_1 = \{0{,}79, 0{,}18, 0{,}03\} \text{ e } P_2 = \{0{,}80, 0{,}16, 0{,}04\},$$

obtemos o mesmo valor de entropia, $H \approx 0{,}600$, mas valores diferentes de $\mathcal{D}$ e, consequentemente, de C. Para $\mathcal{D}$ temos 0,324 e 0,334, e para C, 0,1944 e 0,2004, no primeiro e no segundo caso, respectivamente. Combinando todos os possíveis valores de a e b, podemos construir o plano complexidade-entropia observado na figura 1.2.

1.3 Plano complexidade-entropia

O fato descrito na figura 1.2 serviu de motivação para Rosso *et al.* [30] ao tentar distinguir entre sinais caóticos e estocásticos. Esses pesquisadores notaram que essa distinção é bastante

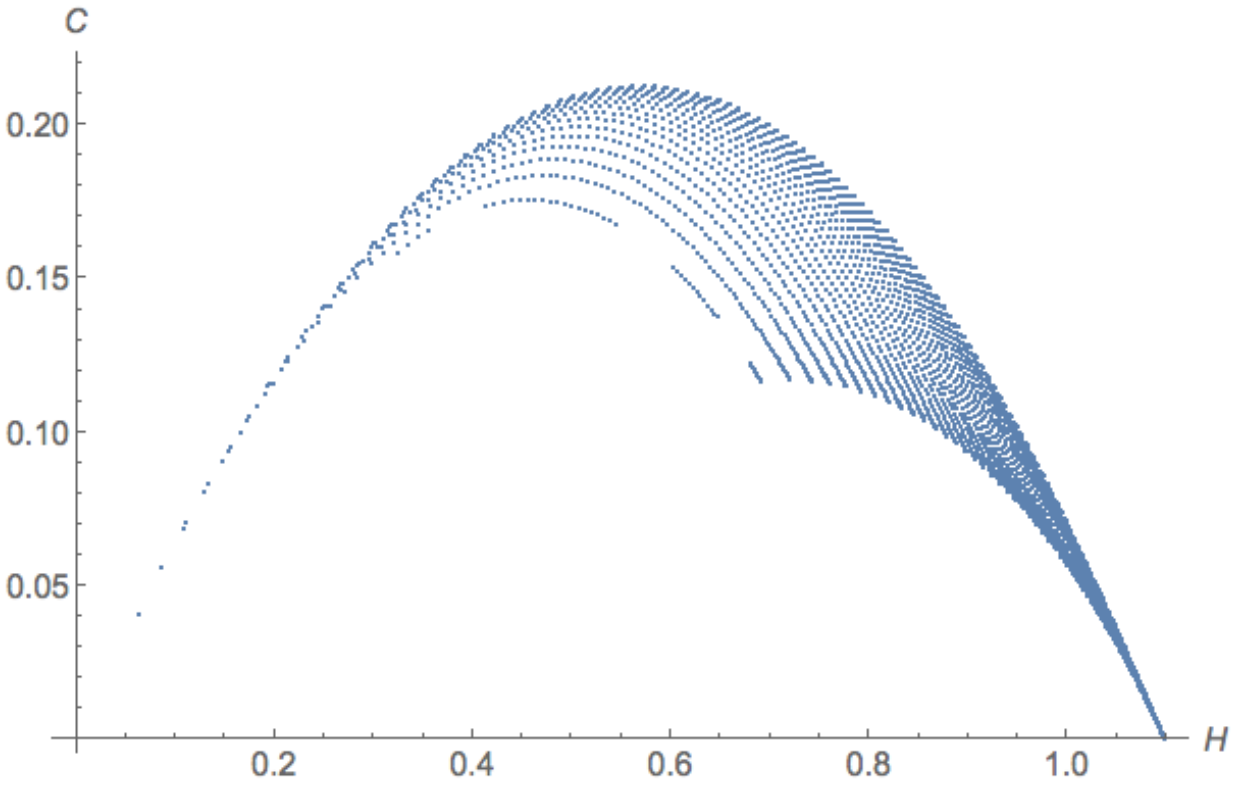

Figura 1.2: Representação do plano complexidade-entropia ilustrando o fato da complexidade não ser uma função unívoca da entropia. Tendo como exemplo o valor de $H \approx 0{,}6$ obtido para as duas distribuições P_1 e P_2 citadas no texto, observamos que existem vários valores possíveis de C ao combinarmos todos os valores de a e b, tais que $a > 0$, $b > 0$ e $(a + b) \leq 1$.

complicada usando apenas a entropia de permutação. Por conta disso, propuseram o uso de uma representação baseada no trabalho de López-Ruiz *et al.* [29], na qual duas medidas são avaliadas: a entropia de permutação normalizada H e a complexidade C associada à divergência de Jensen-Shannon (equação 1.18). Esse diagrama, no qual a entropia é o eixo horizontal e a complexidade o vertical, foi chamado de *complexity-entropy causality plane* (plano complexidade-entropia).

De maneira mais específica, Rosso *et al.* [30] usaram a abordagem de Bandt e Pompe para extrair a distribuição dos padrões ordinais $P = \{p(\pi_i)\}_{i=1,\ldots,d!}$ de várias séries temporais simuladas, cuja dinâmica é conhecida ser caótica ou estocástica. Usando essa distribuição, eles calcularam a entropia de permutação normalizada

$$H[P] = -\frac{1}{\log d!} \sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i), \tag{1.19}$$

e a complexidade estatística

$$C[P] = H[P]\mathcal{D}[P], \tag{1.20}$$

sendo $\mathcal{D}[P]$ o desequilíbrio medido pela divergência de Jensen-Shannon [50], ou seja,

$$\mathcal{D}[P] = \mathcal{D}_0 \left\{ H\left[\frac{P + P_e}{2}\right] - \frac{H[P]}{2} - \frac{H[P_e]}{2} \right\}. \tag{1.21}$$

Note que

$$\frac{P + P_e}{2} = \left\{ \frac{p(\pi_i) + 1/d!}{2} \right\}_{i=1,\ldots,d!} \tag{1.22}$$

e $\mathcal{D}_0$ é uma constante de normalização. Essa constante pode ser obtida encontrando a distribuição P^* que maximiza $\mathcal{D}$ via método dos multiplicadores de Lagrange, similar ao caso de S_{max}. Entretanto, como $\mathcal{D}$ é uma distância entre P e P_e, P^* deve ser uma distribuição com apenas uma componente diferente de zero, isto é, $P^* = \{\delta_{i1}\}_{i=1,\ldots,d!}$. Nesse caso, a constante $\mathcal{D}_0$ fica

$$\mathcal{D}_0 = \left\{ -\frac{1}{2}\left(\frac{d! + 1}{d!} \log(d! + 1) - \log d! - 2\log 2 \right) \right\}^{-1}. \tag{1.23}$$

1.4 Generalização do plano complexidade-entropia para sistemas multidimensionais

Percebendo que as medidas de complexidade descritas anteriormente haviam sido aplicadas apenas em dados unidimensionais, Ribeiro *et al.* [37] propuseram uma extensão para sistemas bidimensionais (ou de dimensão maior) do plano complexidade-entropia.

Seguindo a ideia do início desse capítulo, vamos exemplificar a extensão do método proposta por Ribeiro *et al.* antes de apresentá-la formalmente.

Ao invés de uma série temporal para representar os dados que anteriormente eram unidimensionais, consideramos agora uma matriz bidimensional de tamanho $n_x \times n_y$. Essa matriz pode representar uma imagem, na qual cada elemento corresponde a um *pixel* dessa imagem. Vamos considerar, por simplicidade, a seguinte matriz de tamanho 3×3

$$A = \begin{pmatrix} 3 & 4 & 8 \\ 5 & 6 & 7 \\ 2 & 8 & 9 \end{pmatrix}.$$

Analogamente ao vetor $\vec{s}$ (definido na equação 1.5), que representa partições na série temporal, no caso bidimensional temos matrizes superpostas de tamanho $d_x \times d_y$ ($d_x, d_y \geq 1$). Por exemplo, para $d_x = d_y = 2$, obtemos quatro submatrizes associadas à matriz A da forma

$$A_i = \begin{pmatrix} a_0 & a_1 \\ a_2 & a_3 \end{pmatrix}$$

as quais são:

$$A_1 = \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 4 & 8 \\ 6 & 7 \end{pmatrix},$$
$$A_3 = \begin{pmatrix} 5 & 6 \\ 2 & 8 \end{pmatrix} \quad \text{e} \quad A_4 = \begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix}.$$

Aplicando o método de Bandt e Pompe, ou seja, associando uma sequência de símbolos a cada padrão ordinal das submatrizes A_i, verificamos que A_1 e A_4 são representadas por "0123", pois seus valores estão em ordem crescente: $a_0 \leq a_1 \leq a_2 \leq a_3$. A submatriz A_2, por sua vez, corresponde à sequência "0231", já que $a_0 \leq a_2 \leq a_3 \leq a_1$. Finalmente, A_3 é associada ao padrão "2013", pois $a_2 \leq a_0 \leq a_1 \leq a_3$. De posse das permutações, calculamos as probabilidades de ocorrência $p(\pi_i)$ de cada uma:

$$p(\text{“0123”}) = 0{,}50,$$
$$p(\text{“0231”}) = 0{,}25,$$
$$p(\text{“2013”}) = 0{,}25.$$

Vale notar que de todas as $(d_x d_y)! = 24$ permutações possíveis, apenas as três anteriores surgiram nesse exemplo, as demais são consideradas com probabilidade de ocorrência nula. Usando o conjunto de probabilidades $P = \{p(\pi_i)\}_{i=1,\ldots,(d_x d_y)!}$, podemos determinar as medidas de complexidade estatística de interesse.

Devemos observar que essa extensão do método não é mais definida de modo único, pois, ao invés de ordenarmos os elementos linha por linha para obtermos o padrão ordinal que representa a permutação π_i, poderíamos ordená-los coluna por coluna. Entretanto, o conjunto P das probabilidades não mudaria, apenas os "nomes" das permutações mudariam entre si. Assim, não há perda de generalidade ao assumirmos uma dada ordem para definir as permutações π_i. Vale mencionar, ainda, que o procedimento proposto por Ribeiro *et al.* também pode ser aplicado em estruturas de dimensão maior do que dois e que recupera o caso unidimensional se $n_y = 1$ e $d_y = 1$.

Ilustrada a extensão do método proposto por Ribeiro *et al.* [37], vamos defini-la formalmente. Para isso consideremos uma matriz $\{y_i^j\}_{i=1,\ldots,n_x}^{j=1,\ldots,n_y}$ de tamanho $n_x \times n_y$. As submatrizes (s_x, s_y) de tamanho $d_x \times d_y$ $(d_x, d_y \geq 1)$, que aqui fazem o papel do vetor $(\vec{s})$, são dadas por

$$(s_x, s_y) \mapsto \begin{pmatrix} y_{s_x-(d_x-1)}^{s_y-(d_y-1)} & y_{s_x-(d_x-2)}^{s_y-(d_y-1)} & \cdots & y_{s_x-1}^{s_y-(d_y-1)} & y_{s_x}^{s_y-(d_y-1)} \\ y_{s_x-(d_x-1)}^{s_y-(d_y-2)} & y_{s_x-(d_x-2)}^{s_y-(d_y-2)} & \cdots & y_{s_x-1}^{s_y-(d_y-2)} & y_{s_x}^{s_y-(d_y-2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{s_x-(d_x-1)}^{s_y-1} & y_{s_x-(d_x-2)}^{s_y-1} & \cdots & y_{s_x-1}^{s_y-1} & y_{s_x}^{s_y-1} \\ y_{s_x-(d_x-1)}^{s_y} & y_{s_x-(d_x-2)}^{s_y} & \cdots & y_{s_x-1}^{s_y} & y_{s_x}^{s_y} \end{pmatrix}, \tag{1.24}$$

com $s_x = d_x, d_x+1, \ldots, n_x$ e $s_y = d_y, d_y+1, \ldots, n_y$. Para todas as $(n_x - d_x + 1)(n_y - d_y + 1)$ submatrizes, calculamos as permutações

$$\pi_i = [(r_0, u_0), (r_1, u_0), \ldots, (r_{d_x-1}, u_0), \ldots, (r_0, u_{d_y-1}), (r_1, u_{d_y-1}), \ldots, (r_{d_x-1}), u_{d_y-1})], \quad (1.25)$$

dos símbolos $(0, 1, \ldots, d_x d_y - 1)$ definidas por

$$\begin{aligned} y_{s_x-r_{d_x-1}}^{s_y-u_{d_y-1}} \leq y_{s_x-r_{d-2}}^{s_y-u_{d_y-1}} \leq \ldots \leq y_{s_x-r_1}^{s_y-u_{d_y-1}} \leq y_{s_x-r_0}^{s_y-u_{d_y-1}} \leq \ldots \\ \leq y_{s_x-r_{d_x-1}}^{s_y-u_0} \leq y_{s_x-r_{d-2}}^{s_y-u_0} \leq \ldots \leq y_{s_x-r_1}^{s_y-u_0} \leq y_{s_x-r_0}^{s_y-u_0}. \end{aligned}$$

Lembrando que agora o sistema possui $(d_x d_y)!$ estados acessíveis, para obtermos a distribuição de probabilidades $P = \{p(\pi_i)\}_{i=1,\ldots,d_x d_y!}$, devemos calcular a frequência relativa de cada uma dessas permutações π_i, definida por

$$p(\pi_i) = \frac{\#\{(s_x, s_y) | s_x \leq n_x - d_x + 1 \text{ e } s_y \leq n_y - d_y + 1;\ (s_x, s_y) \text{ é do tipo } \pi_i\}}{(n_x - d_x + 1)(n_y - d_y + 1)}. \quad (1.26)$$

Analogamente ao caso unidimensional, os únicos parâmetros necessários para determinar as probabilidades $p(\pi_i)$ são as *embedding dimensions* d_x e d_y. Essas, devem seguir a relação $(d_x d_y)! \ll n_x n_y$ para obtermos uma estatística confiável.

Finalmente, reescrevendo a entropia de permutação normalizada no caso bidimensional, obtemos

$$H[P] = \frac{S[P]}{S_{max}} = \frac{S[P]}{\log[(d_x d_y)!]}, \quad (1.27)$$

na qual $S[P]$ é a entropia de Shannon da distribuição de probabilidades P (equação 1.13) e $S_{max} = S[P_e] = \log[(d_x d_y)!]$ é obtida no caso da distribuição uniforme P_e.

A complexidade estatística é definida por

$$C[P] = \mathcal{D}[P] H[P], \quad (1.28)$$

sendo o desequilíbrio $\mathcal{D}$ definido em termos da divergência de Jensen-Shannon, ou seja,

$$\mathcal{D}[P] = \mathcal{D}_0 \left\{ H\left[\frac{P + P_e}{2}\right] - \frac{H[P]}{2} - \frac{H[P_e]}{2} \right\}, \quad (1.29)$$

com

$$\mathcal{D}_0 = \left\{ -\frac{1}{2} \left(\frac{(d_x d_y)! + 1}{(d_x d_y)!} \log[(d_x d_y)! + 1] - 2\log[2(d_x d_y)!] + \log[(d_x d_y)!] \right) \right\}^{-1} \quad (1.30)$$

sendo o valor máximo de $\mathcal{D}[P]$, obtido quando uma das componentes da distribuição P é igual a um e todas as outras são nulas.

Assim, encerramos a apresentação da técnica de entropia e complexidade de permutação.

Como o leitor irá notar, essas medidas formam a base teórica de quase todos os sistemas que analisamos nesse trabalho. Em particular, no capítulo 2 utilizamos a versão bidimensional do plano complexidade-entropia para analisar uma grande base de dados de imagens de pinturas [32]. A mesma técnica foi usada no capítulo 3 para extrair propriedades físicas a partir de texturas de cristais líquidos [33]. Por fim, no capítulo 4, empregamos a versão unidimensional da entropia e complexidade de permutação para investigar diversos aspectos sobre a eficiência informacional do mercado de criptomoedas [35] e o comportamento coletivo da eficiência de mercados de ações mundial [36]. Essas aplicações ilustram bem a versatilidade e o potencial dessa abordagem baseada em conceitos de Física Estatística.

1.5 Métodos de aprendizagem estatística

Como já mencionamos na introdução dessa tese, um dos principais objetivos ao estudar sistemas complexos é identificar comportamentos coletivos emergentes, ou seja, propriedades que surgem apenas a partir da interação e relacionamento entre as partes individuais do sistema. Para investigar esses comportamentos coletivos, uma possibilidade é utilizarmos métodos de aprendizagem estatística [31], como regressões e classificações baseadas em aprendizado de máquina (*machine learning*) e agrupamentos hierárquicos.

A ideia de possuir uma máquina que fosse capaz de reproduzir o comportamento inteligente humano define de maneira ampla o campo da inteligência artificial. Citando McCarthy, o primeiro a utilizar esse termo em 1956, definimos inteligência artificial como "a ciência e engenharia de criar máquinas inteligentes, especialmente programas de computador inteligentes" [52] (tradução livre). Nesse contexto, técnicas de aprendizado de máquina podem ser entendidas como um subcampo da inteligência artificial. O primeiro uso do termo *"machine learning"* é atribuído a Samuel [53] em um artigo publicado em 1959, no qual ele verifica a possibilidade de programar um computador para aprender a jogar o jogo de damas. Foi por volta desse mesmo período que Rosenblatt propôs o algoritmo *"perceptron"* [54], que é considerado o primeiro uso de redes neurais artificiais para reconhecimento de padrões e formas.

Apesar desses e de outros desenvolvimentos importantes, foi somente após o início do século XXI que os métodos de aprendizado de máquina e, particularmente, a classe de métodos de aprendizado profundo (*deep learning*) [55–57] começaram a ser amplamente empregados em várias áreas para lidar com as mais distintas questões, que incluem auxiliar em diagnósticos médicos de diversas doenças [58–65], prever o preço de ações da bolsa de valores [66], implementar tarefas de processamento de linguagem natural [67] e de visão computacional [57,68,69] e até recomendar possíveis amigos que você conheça mas ainda não esteja conectado em uma rede social [70]. Essa ampla gama de aplicações práticas associada ao desenvolvimento de novas técnicas e algoritmos, fez com que esses métodos baseados

em *machine learning* experimentassem um crescimento vertiginoso em sua popularidade nas últimas décadas [71].

De maneira simplificada, ao implementar esses métodos de aprendizagem estatística, a máquina "aprende" a partir de dados, ou seja, melhora progressivamente seu desempenho em uma tarefa específica, sem que para isso ela seja diretamente programada. Essas tarefas geralmente estão ligadas à previsões quando, por exemplo, queremos prever valores futuros para uma ação baseado em valores passados, ou inferências, quando ao invés de prevermos os valores, estamos interessados em entender como o preço varia em função de outras características. Uma das grandes vantagens de abordar problemas por meio de métodos baseados em *machine learning* é a capacidade que eles possuem para extrair características importantes a partir de conjuntos de dados desordenados e de larga escala. Tarefas desse tipo costumavam ser realizadas manualmente por especialistas, demandando muito tempo e esforço. Além disso, em muitos casos uma abordagem manual é impossível de ser realizada devido à complexidade e tamanho do conjunto de dados.

Existem duas categorias principais nas quais os problemas de aprendizagem estatística podem ser classificados: aprendizagem supervisionada e não supervisionada [31]. Na aprendizagem supervisionada, o algoritmo aprende a partir de um conjunto de dados chamado conjunto de treino, o qual conta com as variáveis de entrada x_i e as variáveis resposta y_i. O objetivo, nesse caso, é estimar a função $\hat{f}$ que mapeia as variáveis de entrada nas variáveis resposta. Para isso, é como se o procedimento estivesse sendo supervisionado, pois como sabemos as variáveis resposta, se o algoritmo faz uma previsão errada no conjunto de treino, ele pode ser corrigido e ter seu desempenho melhorado a cada iteração. Na aprendizagem não supervisionada, por sua vez, temos apenas o conjunto de dados de entrada e não conhecemos as variáveis resposta e, portanto, não podemos treinar um algoritmo. Nesse caso, o objetivo é encontrar possíveis relações e padrões entre os dados, como estruturas de agrupamento hierárquico de acordo com similaridades extraídas pelo algoritmo.

Uma outra classificação geral das tarefas de aprendizagem estatística supervisionadas diz respeito aos tipos de variáveis resposta, que podem ser quantitativas ou qualitativas [31]. Em geral, variáveis quantitativas assumem valores numéricos reais, enquanto as qualitativas descrevem classes e, por isso, também são denominadas por variáveis categóricas. Os problemas de *machine learning* em que a variável resposta é quantitativa são conhecidos por tarefas de *regressão*. Quando a variável resposta é qualitativa, temos uma tarefa de *classificação*.

Para ilustrar essas diferentes classificações, tomemos como exemplos os estudos que serão apresentados no capítulo 3. Conforme veremos, na primeira seção propomos uma abordagem para extrair propriedades físicas de cristais líquidos a partir da entropia e complexidade de permutação calculadas para texturas desses materiais utilizando regressores e classificadores. Por exemplo, ao prever a temperatura da amostra (uma grandeza contínua), realizamos uma tarefa de regressão. Por outro lado, para prever o passo de um cristal líquido coles-

térico dentre um conjunto discreto de passos, efetuamos uma tarefa de classificação. No estudo apresentado na segunda seção do capítulo 3, empregamos redes convolucionais neurais profundas para realizar praticamente as mesmas tarefas de regressão e classificação de propriedades de cristais líquidos, porém sem a necessidade de extrair características (entropia e complexidade de permutação) das texturas desses materiais. Percebemos, portanto, que existe um grande número de métodos de aprendizagem estatística e a decisão sobre qual deles empregar em um determinado problema é importante e muitas vezes empírica, visto que não há uma regra geral para essa escolha ou métodos que produzam bons resultados independentemente do conjunto de dados analisado.

Sendo assim, precisamos aferir o desempenho desses diferentes métodos de alguma maneira para encontrar aquele com a maior precisão para o problema em questão. Para isso, uma medida frequentemente avaliada é o erro quadrático médio EQM, definido por

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{1.31}$$

com $\hat{f}$ representando uma estimativa da função que mapeia as variáveis de entrada nas variáveis resposta, $\hat{f}(x_i)$ o valor previsto por $\hat{f}$ e y_i o valor verdadeiro para a i-ésima observação. Por definição, essa quantidade será cada vez menor quanto mais próximas das respostas verdadeiras forem as respostas previstas. Para um dado valor x_0, o erro quadrático médio pode ser decomposto na soma de três quantidades: a variância de $\hat{f}(x_0)$, o quadrado do viés de $\hat{f}(x_0)$ e a variância dos termos de erro (também chamada de erro irredutível). Matematicamente, escrevemos

$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Viés}(\hat{f}(x_0)]^2 + \mathrm{Var}(\epsilon), \tag{1.32}$$

na qual $E(y_0 - \hat{f}(x_0))^2$ representa o valor esperado para o erro quadrático médio, que pode ser obtido ao estimar $\hat{f}$ repetidamente utilizando vários conjuntos de treino e testando cada função estimada em x_0. O valor esperado geral para o erro quadrático médio é calculado pela média de $E(y_0 - \hat{f}(x_0))^2$ para todos os possíveis valores de x_0. Uma vez que a variância e o quadrado do viés de $\hat{f}(x_0)$ são positivos por definição, o limite inferior para o valor esperado do erro quadrático médio é dado pelo erro irredutível. Em vista disso, para encontrar o valor mínimo dessa quantidade, devemos minimizar tanto o termo da variância quanto o termo do viés.

Porém, essa tarefa não é tão simples, visto que essas quantidades estão intimamente relacionadas entre si e com a complexidade do método em questão, dando origem ao fenômeno chamado *trade-off* entre o viés e a variância [71]. Os erros decorrentes de variância estão relacionados à sensibilidade na estimativa de $\hat{f}$ em resposta a pequenas flutuações no conjunto de treino, ou seja, se um método possui variância alta, pequenas mudanças no conjunto de

treino resultam em estimativas muito diferentes para a função $\hat{f}$. Geralmente, modelos ou algoritmos mais complexos costumam apresentar maior variância. Os erros decorrentes do viés, por sua vez, surgem devido às hipóteses simplificadas assumidas pelo método, no sentido em que ao aproximarmos um problema complexo por um modelo muito mais simples, o viés na estimativa de $\hat{f}$ usualmente será alto. Os erros de viés são definidos pela diferença entre o valor esperado predito pelo método e o valor correto. Modelos mais complexos usualmente possuem um viés menor, ou seja, tendem a capturar melhor as relações relevantes do conjunto de dados, no sentido em que são mais flexíveis e podem gerar várias formas para $\hat{f}$.

Ainda com relação à complexidade do modelo, enfatizamos que os erros relacionados à variância são mais proeminentes quando o método é complexo ao ponto de modelar quase exatamente o conjunto de treino, inclusive o ruído, mas falhar em prever observações futuras utilizando dados ainda não apresentados ao algoritmo. Esse fato é conhecido por *overfitting* [31]. Por outro lado, os erros relacionados ao viés se destacam quando o método não é complexo o suficiente para capturar os padrões existentes no conjunto de dados, fato conhecido por *underfitting* [31]. Portanto, existe um *trade-off* entre minimizar o viés e a variância com relação a complexidade do método, de modo que a complexidade ideal é aquela que evita ao máximo tanto a ocorrência de *overfitting* quanto de *underfitting*.

Na prática, é comum utilizar métodos de reamostragem para avaliar esse *trade-off* e determinar o melhor modelo com seus parâmetros ótimos que minimizam o erro. Dentre os principais métodos com essa finalidade, temos a abordagem de validação cruzada [31] de n camadas. De modo geral, esses métodos de reamostragem utilizam amostras do conjunto de dados e ajustam o modelo para cada uma, a fim de obter mais informações sobre o método. No caso do método de validação cruzada de n camadas, primeiro dividimos aleatoriamente o conjunto de dados em n partes com aproximadamente o mesmo tamanho. Em seguida, uma das partes é separada como conjunto de teste e as $(n-1)$ partes restantes são utilizadas para treinar o algoritmo. Esse procedimento é repetido até que todas as partes tenham sido utilizadas como conjunto de teste. A precisão obtida a partir do conjunto de treino é o *score* de treino e a precisão obtida a partir do conjunto de teste é o *score* de validação cruzada ou *score* de teste. Ao fim, calculamos médias e intervalos de confiança para essas quantidades usando as n estimativas obtidas. Gráficos desses *scores* em função dos parâmetros do modelo são conhecidos por curvas de validação. Por outro lado, as curvas de aprendizagem indicam a dependência desses *scores* com o tamanho do conjunto de treino. Essa última questão também é relevante, visto que conjuntos de treino muito pequenos podem não ser suficientes para ajustar corretamente o modelo, enquanto dados que não sejam necessários podem introduzir ruído ao modelo. Tanto as curvas de validação quando as de aprendizagem são ferramentas importantes para avaliar a qualidade do modelo e, por isso, são estimadas em análises nas quais aplicamos métodos de aprendizagem estatística.

Para ilustrar o procedimento de um desses métodos de aprendizagem estatística, vamos

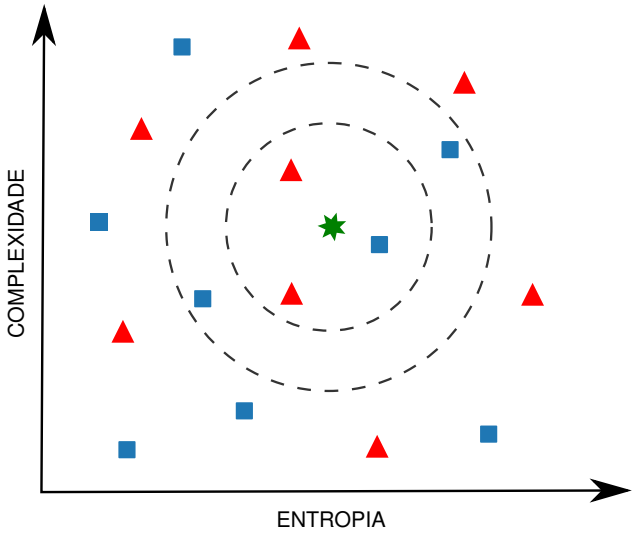

Figura 1.3: Exemplo de aplicação do algoritmo k-vizinhos mais próximos em uma tarefa de classificação. A classe da observação representada no centro da figura é determinada com base na classe mais frequente dos k-vizinhos mais próximos. Para k=3, nos limitamos a observar os três primeiros vizinhos, que são delimitados pela circunferência tracejada interna. Nesse caso, temos dois triângulos e um quadrado, portanto, a nova observação seria um triângulo. Já se k for igual a 5, devemos analisar para os 5 primeiros vizinhos delimitados pela circunferência tracejada externa. Nesse caso, temos três quadrados e dois triângulos, e a nova observação seria classificada como um quadrado.

apresentar o algoritmo conhecido por k-vizinhos mais próximos [72]. Esse método pode ser aplicado em problemas de regressão e classificação e, além de ser um dos métodos de aprendizagem estatística mais simples, uma de suas vantagens é possuir apenas um parâmetro: o número de vizinhos k. Na figura 1.3, mostramos um exemplo da aplicação de um algoritmo classificador de k-vizinhos mais próximos para prever a classe de uma nova observação baseada na classe mais frequente dos k-vizinhos mais próximos. A nova observação, cuja classe é desconhecida a priori, está representada no centro da figura. Para k=3, nos limitamos a observar os três primeiros vizinhos, que são delimitados pela circunferência tracejada interna. Nesse caso, temos dois triângulos e um quadrado; portanto, a nova observação é classificada como um triângulo. Para k=5, analisamos os 5 primeiros vizinhos delimitados pela circunferência tracejada externa. Assim, temos três quadrados e dois triângulos e a nova observação é classificada como um quadrado. Os algoritmos de aprendizagem estatística utilizados nesse trabalho são implementados utilizando a linguagem de programação *Python* em conjunto com as bibliotecas *Numpy* [73], *SciPy* [74] e *scikit-learn* [75], todas disponíveis livremente com seus códigos-fonte.

1.6 Redes convolucionais neurais

Apesar dos grandes e importantes avanços em várias aplicações de algoritmos de aprendizado de máquina, o processo de extrair informações significativas de imagens, ou seja, replicar a função do sistema visual humano, tem se mostrado uma tarefa mais desafiadora [69, 76]. Redes convolucionais neurais são consideradas ferramentas do estado-da-arte para analisar imagens e possuem a vantagem de não requererem a extração de características das imagens de maneira manual. Em particular, essas redes neurais profundas usam uma cascata hierárquica de convoluções e funções não-lineares que aprendem automaticamente representações e características de baixo nível diretamente das imagens de entrada [77]. Esse é um dos motivos dessas redes neurais convolucionais profundas serem muito boas em identificar objetos em imagens.

Na verdade, um exemplo notório do sucesso dos algoritmos de aprendizagem profunda (*deep learning*) é documentado no desafio de reconhecimento visual em grande escala *ImageNet* [78], uma competição anual entre algoritmos para detecção e classificação de objetos. A introdução de um modelo de rede neural profunda (AlexNet) proposto por Krizhevsky *et al.* [79] em 2012 é considerada o maior avanço na competição não somente porque a taxa de erro foi reduzida de 26% para 16, 4%, mas principalmente devido ao fato desses algoritmos terem se tornado os principais competidores desde então [78]. Também foi um algoritmo de aprendizagem profunda (ResNet) o primeiro a ultrapassar o nível de desempenho humano no conjunto de imagens do desafio *ImageNet* em 2015 [80, 81].

Iremos introduzir de maneira breve os conceitos fundamentais das redes convolucionais neurais [55–57] que, em nosso caso, serão empregadas em um dos estudos apresentados no capítulo 3 para prever propriedades físicas de cristais líquidos diretamente das imagens de suas texturas. Essas redes são um tipo particular de rede neural artificial cuja unidade básica é um neurônio ou um nó. O *design* das redes neurais artificiais é inspirado em conceitos de redes neurais biológicas e consiste em camadas de neurônios que estão completamente conectados uns aos outros de maneira ponderada. Cada neurônio recebe valores de entrada da camada anterior, calcula a soma ponderada desses valores, adiciona um termo de viés, avalia uma função não-linear (função de ativação), e envia o valor obtido por essa função para a próxima camada. Esse processo imita de maneira simples o comportamento de neurônios biológicos que disparam sob estímulos suficientes. O processo de treinar uma rede neural artificial consiste em ajustar os pesos e os termos de viés de maneira que os sinais de entrada produzam valores de saída que coincidam ou aproximem-se dos valores fornecidos pelo conjunto de treino. Os valores dos pesos e dos vieses são atualizados ao avaliar uma função de perda que quantifica o erro de saída em um processo conhecido como *"backpropagation."* Durante esse processo, um algoritmo de gradiente descendente estocástico é usado para atualizar de maneira iterativa os pesos e os vieses a fim de minimizar a função de perda. Nesse processo

de ajuste ou otimização, cada iteração completa dos dados de treinamento caracteriza uma época.

A principal diferença entre redes neurais usuais e redes neurais convolucionais é a existência de camadas convolucionais. Diferentemente das camadas completamente conectadas, os "neurônios" em camadas convolucionais recebem os dados de entrada de regiões espacialmente pequenas e contínuas da camada anterior. Esses dados de entrada em janelas são multiplicados por filtros que possuem os mesmos pesos para todo o conjunto de entrada. Dessa maneira, redes convolucionais preservam a estrutura espacial e otimizam os pesos dos filtros que são os responsáveis por detectar e extrair características de baixo nível em localizações diferentes dos dados de entrada (geralmente imagens). Para definir uma camada convolucional, precisamos especificar o tamanho das janelas espaciais (tamanho do filtro) e a sobreposição entre janelas adjacentes (*stride*). Por exemplo, um filtro de tamanho 2×2 e *stride* ou passo igual a 1 opera sobre janelas com dimensões de 2×2 *pixels* (se a entrada for uma imagem), movendo-se em passos unitários sobre os dados de entrada. Além das camadas convolucionais, essas redes geralmente possuem camadas de agrupamento (*pooling layers*) e de subamostragem (*downsampling layers*). Uma camada de agrupamento opera de maneira similar a uma camada convolucional, mas ao invés de calcular a soma ponderada, essa camada realiza cálculos simples para cada região, tais como o valor máximo (*max pooling*) ou valor médio (*average pooling*) dentro de cada janela. Essas camadas de agrupamento sintetizam a presença de características, tornando suas representações invariantes a pequenas translações nos dados de entrada. Além disso, essas camadas reduzem a dimensão dos dados e, consequentemente, a quantidade de parâmetros, o que por sua vez, melhora a eficiência computacional do modelo. As redes neurais convolucionais usadas ao longo desse trabalho foram implementadas usando as bibliotecas *TensorFlow* [82] e *Keras* [83].

CAPÍTULO 2

Quantificando conceitos e aspectos da história da arte

Neste capítulo, apresentamos um estudo em grande escala de aproximadamente 140 mil imagens de obras de arte que abrangem quase um milênio da história da arte [32]. Baseados nos padrões espaciais locais das imagens digitais dessas obras, calculamos a entropia e a complexidade de permutação de cada uma delas. Mostramos que essas medidas mapeiam o grau de ordem visual dessas obras em uma escala de ordem/desordem e simplicidade/complexidade, as quais refletem bem categorias qualitativas propostas por historiadores da arte. Além disso, o comportamento dinâmico dessas medidas revela uma clara evolução temporal da arte, marcada por transições que coincidem com os principais períodos da história da arte. Nossa pesquisa mostra ainda que estilos artísticos distintos possuem diferentes valores médios de entropia e complexidade, permitindo uma organização hierárquica e o agrupamento desses estilos. Mostramos também que métodos de aprendizagem estatística baseados apenas nessas medidas de complexidade podem classificar obras de arte com respeito ao seu estilo de uma maneira rápida e eficaz.

2.1 Introdução e apresentação dos dados

Abordagens inspiradas em Física têm sido aplicadas com sucesso em uma grande variedade de disciplinas, incluindo sistemas econômicos e sociais [7, 84, 85]. O impacto e a popularidade dessas pesquisas cresceu vertiginosamente nos últimos anos. Em grande parte, esse fato reflete a quantidade enorme de informação digital disponível sobre os mais diversos assuntos e em um grau de detalhe impressionante. Conforme já discutimos, esses dados têm permitido que pesquisadores estudem, de forma quantitativa, uma variedade de sistemas complexos com uma riqueza de detalhes inimaginável há uma década atrás. A caracterização

em grande escala de artes visuais estaria entre esses sistemas, não somente devido à falta de dados, mas também porque o estudo da arte costuma ser intrinsecamente qualitativo. Entretanto, abordagens quantitativas visando caracterizar o domínio das artes visuais podem contribuir muito para um melhor entendimento da evolução cultural humana, assim como para questões mais práticas, como a caracterização e classificação de imagens.

Embora a escala de alguns estudos recentes tenha mudado drasticamente, o uso de técnicas quantitativas no estudo da arte possui algum precedente. Os esforços nesse sentido datam do ano de 1933, quando o matemático americano Birkhoff publica seu livro intitulado *"Aesthetic Measure"* [86], no qual uma medida quantitativa de estética é definida como a razão entre ordem (número de regularidades encontrada em uma imagem) e complexidade (número de elementos em uma imagem). No entanto, a aplicação de técnicas quantitativas na caracterização de obras de arte é muito mais recente. Entre os trabalhos pioneiros, temos o artigo de Taylor *et al.* [87], no qual as pinturas de Jackson Pollock são caracterizadas por uma dimensão fractal que aumenta no decorrer da carreira artística desse importante pintor do movimento Expressionismo Abstrato. Esse artigo pode ser considerado um marco para o estudo quantitativo das artes visuais, inspirando muitas outras aplicações relacionadas à determinação da autenticidade de pinturas [88–92], evolução de artistas específicos [93, 94], propriedades estatísticas de pinturas particulares [95] e de artistas [96–98], movimentos artísticos [99] e muitas outras formas de expressão visual [100–102]. Os avanços mais recentes desse campo de pesquisa emergente e em rápido crescimento são documentados de forma abrangente em vários *proceedings* de conferências e edições especiais de revistas científicas [103–105], cujas contribuições também envolvem ferramentas para restauração de obras de arte, problemas de autenticidade e procedimentos para avaliação do estilo artístico.

Apesar do grande interesse nesse tema, poucos trabalhos foram dedicados ao estudo em grande escala de obras de arte levando em conta uma perspectiva histórica. Em 2014, Kim *et al.* [106] analisaram 29 mil imagens e mostraram que a distribuição do uso das cores é bastante distinta entre os períodos históricos da pintura ocidental. Além disso, os mesmos autores verificaram que o expoente de rugosidade associado à representação em escala de cinza dessas pinturas apresenta uma tendência de crescimento ao longo dos anos. Em um trabalho mais recente, Lee *et al.* [107] analisaram aproximadamente 180 mil obras de arte com foco na evolução do contraste de cores. Entre outros resultados, os autores observaram um crescimento repentino na diversidade de contraste de cores após o ano 1850, além de mostrarem que essa quantidade pode ser utilizada para capturar informações sobre estilos artísticos.

No entanto, exceto pela introdução do expoente de rugosidade, as pesquisas anteriores se concentraram predominantemente, na evolução dos perfis de cores. De fato, os padrões espaciais dessas imagens, os quais representam um aspecto fundamental das obras de arte, permanecem pouco compreendidos. Nesse sentido, a principal contribuição de nosso traba-

lho [32] é suprir um pouco da escassez de estudos nessa direção.

Para investigar esses padrões espaciais, construímos uma grande base de dados composta de 137.364 imagens digitais de obras de arte a partir da enciclopédia virtual de artes visuais WikiArt.org (www.wikiart.org). Fizemos o *download* automatizado dessas imagens e de vários metadados relacionados a cada obra, como artista (são 2.391 artistas diferentes), data de produção e estilo artístico (Impressionismo, Surrealismo, Barroco, etc). Para analisar a evolução temporal, excluímos todas as imagens cujas datas de produção da obra não estavam especificadas (33.724 imagens). A figura 2.1A mostra o número de imagens por ano em nosso conjunto de dados em escala log-linear. Notamos que essas obras de arte foram criadas entre os anos 1031 e 2016, abrangendo quase um milênio da história da arte. A figura 2.1B mostra que a fração acumulada de obras em nosso conjunto de dados é bem aproximada por um crescimento exponencial com tempo característico igual a $\tau = 111 \pm 1$ anos. Consequentemente, o número acumulado de obras de arte tem dobrado a cada 77 anos. Além disso, mais de 50% dessas obras foram produzidas após a primeira década do século XX, um período marcado pelo desenvolvimento de uma grande variedade de movimentos artísticos.

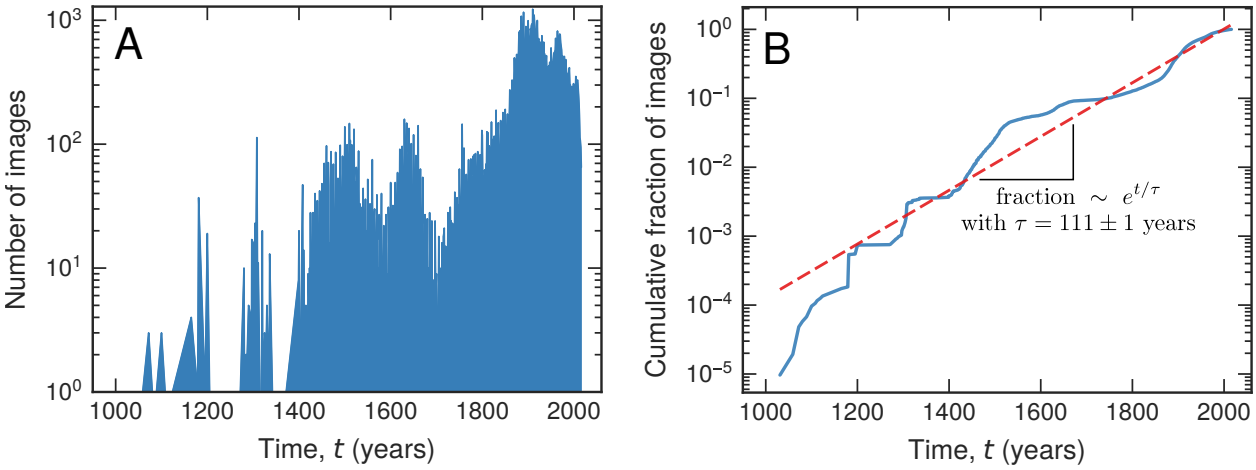

Figura 2.1: Distribuição das imagens das obras de arte ao longo dos anos. (A) Número de imagens por ano em nosso conjunto de dados em escala log-linear. (B) Fração acumulada das obras de arte ao longo dos anos (curva azul) em escala log-linear. Essa fração $[f(t)]$ é aproximada por um crescimento exponencial $[f(t) \propto \exp(t/\tau)]$ com tempo característico igual a $\tau = 111 \pm 1$ anos (linha tracejada). Observamos que a maioria das obras de arte foram produzidas após o início do século XX, sendo mais de 50% delas criadas após o ano 1912.

2.2 Representação matricial das imagens das obras de arte

Os arquivos das imagens digitais das obras de arte estão no formato JPEG com 24 *bits* por *pixel*, sendo 8 *bits* para cada uma das três cores no espaço de cores RGB. Isso significa que cada *pixel* da imagem é caracterizado por uma entre 256 possíveis intensidades de vermelho (R), verde (G) e azul (B), permitindo um total de $256^3 = 16.777.216$ variações de cores.

Do ponto de vista computacional, uma imagem pode ser representada por uma matriz de três camadas com dimensões n_x (a largura da imagem) por n_y (a altura da imagem), na qual as camadas correspondem a cada uma das três cores do espaço RGB e cujos elementos (variando de 0 a 255) representam a intensidade da cor. Para nossas análises, calculamos a soma das intensidades das três cores para cada *pixel*, de modo que cada imagem é representada por uma matriz simples. A partir dessa matriz calculamos a entropia H e a complexidade de permutação C, seguindo os procedimentos descritos na seção 1.4. A figura 2.2 ilustra esse procedimento.

Essa abordagem é similar a transformação usual de uma imagem colorida para escala de cinza, exceto pelo fato de que, nesse procedimento, utilizamos a média ponderada dos valores das três camadas. Uma das combinações mais utilizadas define a chamada reflectância ou luminância [108] e corresponde a calcular $0{,}2125R + 0{,}7154G + 0{,}0721B$, sendo R, G e B as intensidades de vermelho, verde e azul, respectivamente. Esses valores para os pesos geralmente são escolhidos para imitar a sensibilidade às cores do olho humano. Nossos resultados são bem robustos ao considerar diferentes escolhas para esses valores, por exemplo, o coeficiente de correlação linear de Pearson entre os valores de H calculados via transformação usual em escala de cinza e utilizando a média simples é 0,989. Esse mesmo coeficiente é igual a 0,992 para os valores de C. Esses valores próximos de 1 indicam que essas diferentes transformações levam a valores muito parecidos para H e C, conforme mostra a figura 2.3.

Em nosso caso, ao calcular H e C utilizamos $d_x = d_y = 2$ como escolha para as *embedding dimensions*, as quais, conforme já vimos, são os únicos parâmetros desse método. Essa escolha se deve ao fato dos valores médios da largura n_x e da altura n_y da imagem serem próximos a 900 *pixels*, como mostra a figura 2.4. Esse resultado praticamente limita nossa escolha a $d_x = d_y = 2$ para que a condição $(d_x d_y)! \ll n_x n_y$ seja satisfeita, conduzindo a uma estimativa confiável para a distribuição de probabilidades dos padrões ordinais.

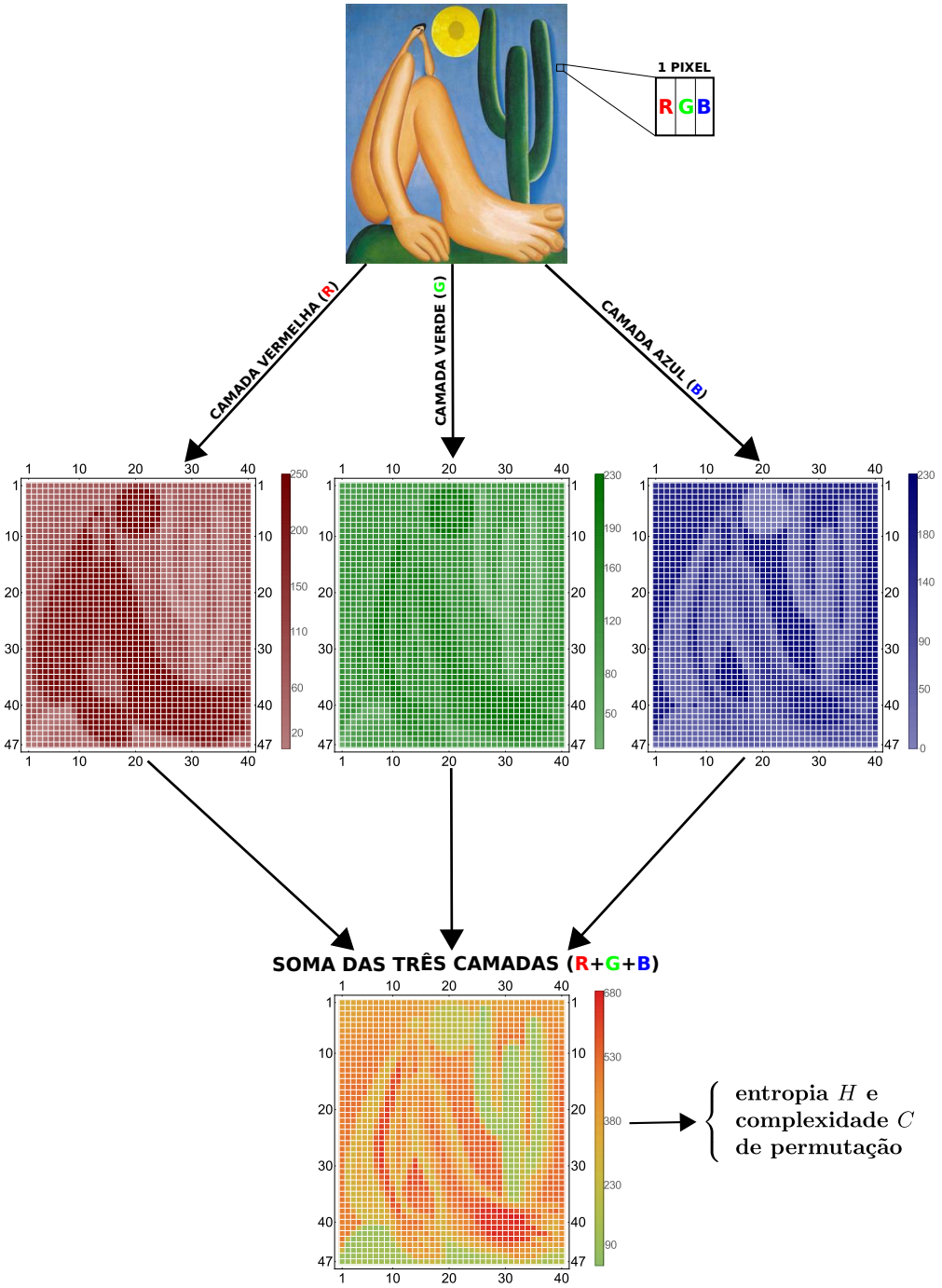

Figura 2.2: Ilustração do procedimento utilizado para processar as imagens e calcular a entropia e a complexidade de permutação. Para o computador, a imagem pode ser interpretada como uma matriz de *pixels* subdividida em três camadas na representação RGB: vermelha (R), verde (G) e azul (B). A imagem mostrada aqui, *Abaporu* de Tarsila do Amaral, possui tamanho 40×47 *pixels* (usamos uma imagem reduzida apenas para fins ilustrativos). A intensidade de cada *pixel* nas camadas R, G e B varia entre 0 e 255, conforme mostram os painéis intermediários. O painel mais abaixo mostra a matriz da soma das três camadas, a partir da qual calculamos os valores de entropia H e complexidade C de permutação.

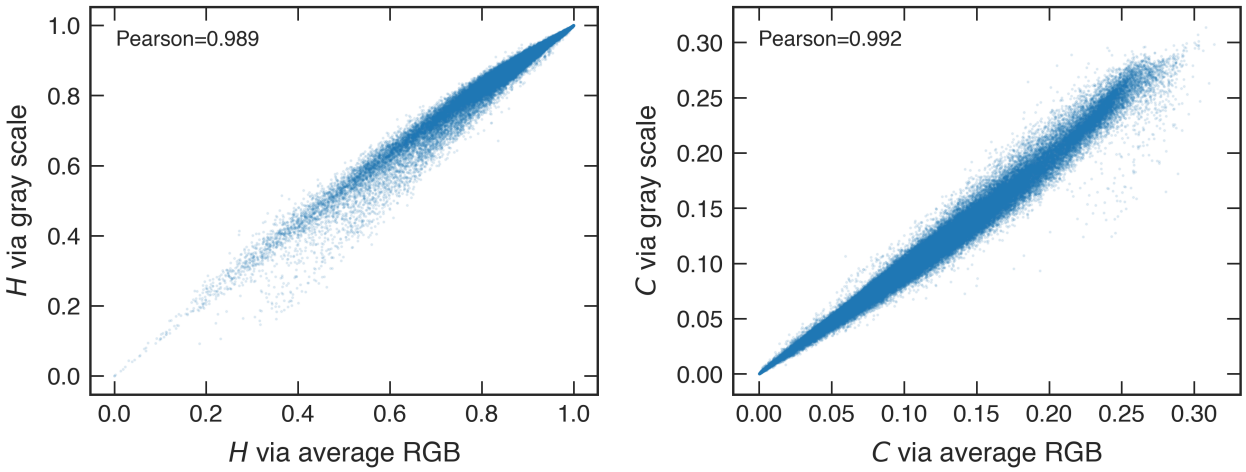

Figura 2.3: Relação entre os valores de H (à esquerda) e C (à direita) calculados pela soma dos canais RGB versus as mesmas quantidades obtidas pela transformação em escala de cinza (luminância). Cada ponto nos gráficos de dispersão mostra os valores de H e C para uma imagem obtidos por meio da soma das intensidades das três cores em cada *pixel* versus as mesmas quantidades calculadas por meio da transformação em escala de cinza (luminância). Observamos que ambas as transformações resultam em valores de H e C fortemente correlacionados, como indicado pelos valores do coeficiente de correlação de Pearson mostrados na figura.

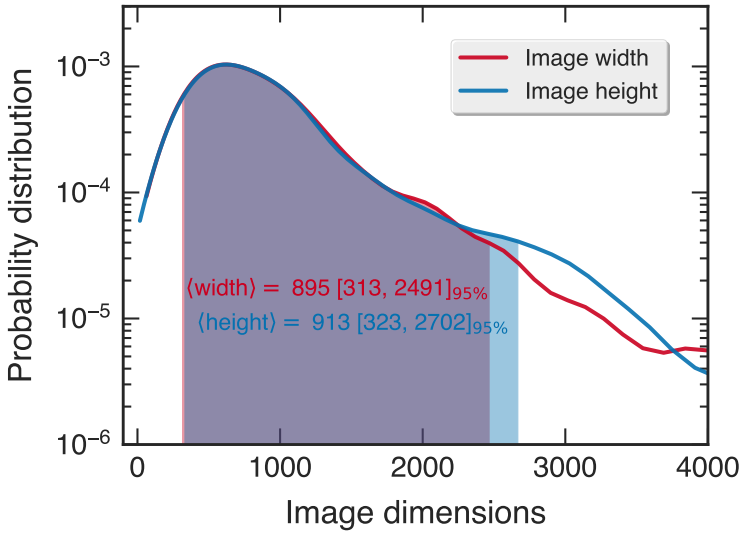

Figura 2.4: Distribuição de probabilidade das dimensões das imagens. As curvas azul e vermelha mostram as distribuições de probabilidade da largura e da altura para todas as imagens em nosso conjunto de dados em escala log-linear. Note que a largura e altura possuem distribuição similar e praticamente o mesmo valor médio (895 *pixels* para a largura e 913 *pixels* para a altura). As regiões sombreadas representam os intervalos de largura e altura que contêm 95% de todas imagens.

2.3 Independência dos valores de H e C com as dimensões das imagens

Conforme vimos na figura 2.4, as imagens digitais obtidas em nossa base de dados não possuem as mesmas dimensões. Essa mesma figura mostra que tanto a largura quanto a altura possuem uma distribuição similar, com valores médios iguais a 895 *pixels* para a largura e 913 *pixels* para a altura. Além disso, 95% das imagens possuem a largura variando entre 313 e 2.491 *pixels* e a altura entre 323 e 2.702 *pixels*.

Por causa dessa grande variação nas dimensões, investigamos se os valores de H e C possuem algum viés devido ao tamanho da imagem. Essa questão é importante já que esperamos que os valores de H e C reflitam os padrões ordinais das imagens e não suas dimensões. Na figura 2.5, mostramos vários gráficos de dispersão dos valores de H e C versus a raiz quadrada da área das imagens ($\sqrt{n_x n_y}$) em escalas diferentes. Visualmente não é possível identificar nenhuma relação e, de fato, o coeficiente de correlação linear de Pearson é muito baixo ($\approx 0,05$) para ambas as relações. Também estimamos o coeficiente de máxima informação (MIC) [109], uma medida não paramétrica que quantifica o grau de associação entre duas variáveis mesmo se elas estiverem correlacionadas de uma maneira não linear. O valor do MIC também é muito pequeno ($\approx 0,07$) para ambas as relações. Portanto, concluímos que os valores de H e C não são afetados pelas dimensões das imagens.

Sendo assim, podemos nos concentrar no problema de quantificar conceitos canônicos da história da arte empregando as medidas de complexidade que calculamos para as imagens das obras de arte.

2.4 Evolução da arte

A comparação cuidadosa entre diferentes obras de arte é um dos principais métodos utilizados por historiadores da arte para entender se e como a arte evoluiu ao longo do tempo. Os trabalhos de Heinrich Wölfflin [110] e Alois Riegl [111], por exemplo, podem ser considerados fundamentais nesse sentido. Eles propuseram a distinção de obras de arte de períodos diferentes por meio de poucas categorias visuais e descritores qualitativos. A comparação visual é, sem dúvida, uma ferramente útil para avaliar estilos artísticos. No entanto, é impraticável aplicar essa abordagem em grande escala. Nessa condição, os métodos computacionais têm a oportunidade de mostrarem suas vantagens. Contudo, qualquer tentativa de quantificar uma obra de arte precisa ser facilmente interpretada em termos de categorias familiares e academicamente relevantes para que possa ser útil.

Com relação a isso, notamos que o plano complexidade-entropia reflete, ao menos em parte, as concepções duais de Wölfflin e a dicotomia de Riegl. De acordo com Wölfflin, é possível classificar uma obra de arte utilizando um conjunto pequeno de pares de caracte-

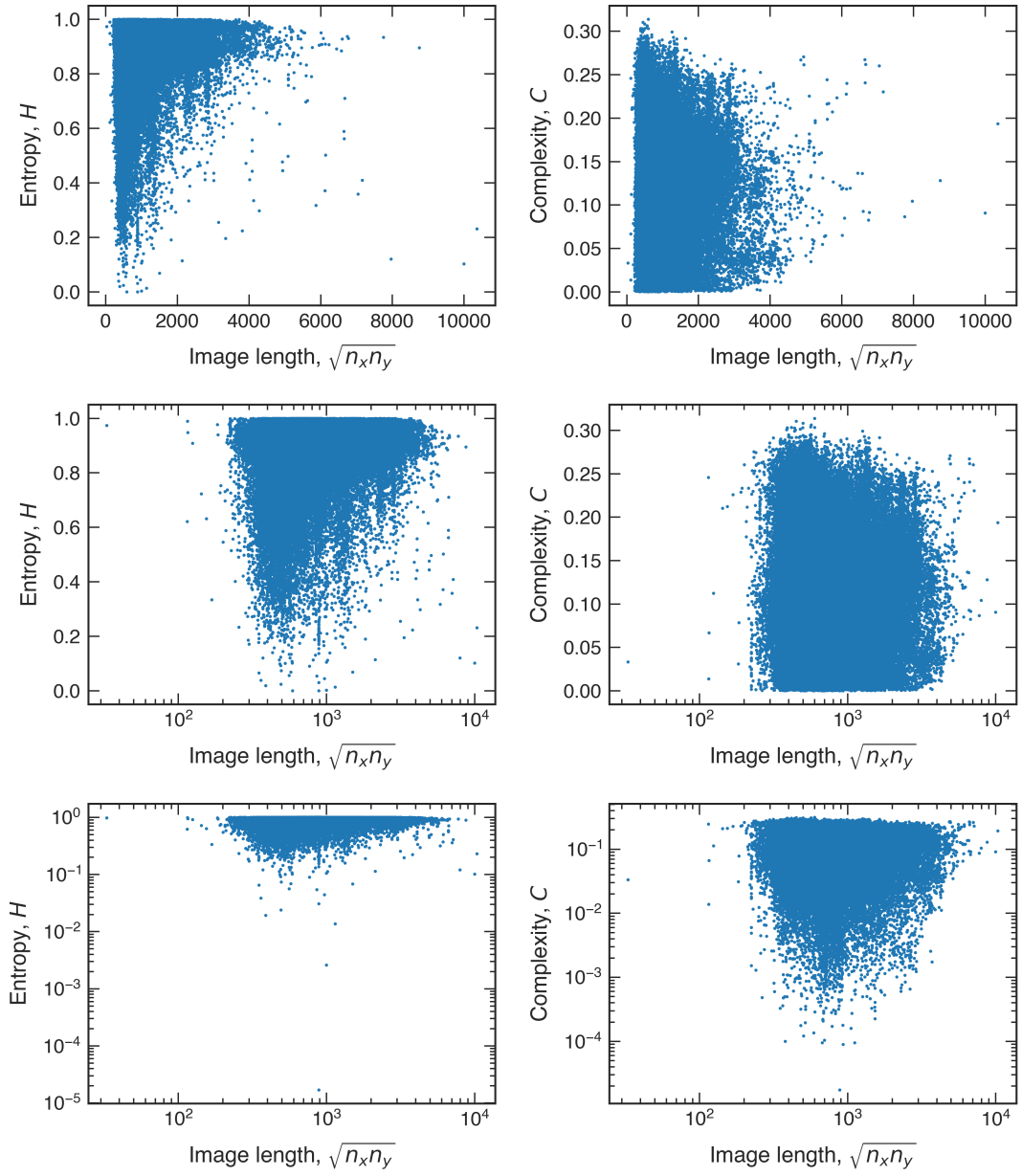

Figura 2.5: As medidas de complexidade H e C não são correlacionadas com as dimensões das imagens. Os gráficos de dispersão mostram os valores de H (painéis à esquerda) e C (painéis à direita) versus o "comprimento da imagem," definido como a raiz quadrada da área da imagem (isto é, $\sqrt{n_x n_y}$, com n_x e n_y, representando a largura e a altura da imagem, respectivamente). A primeira fileira horizontal de gráficos mostra a relação em escala linear, a segunda em escala linear-log e a terceira em escala log-log. Cada ponto representa uma imagem em nosso conjunto de dados. Não observamos correlação entre as medidas de complexidade e o comprimento da imagem. Em particular, o coeficiente de correlação linear de Pearson é $\approx 0{,}05$ para a relação entre o comprimento da imagem e H e $\approx 0{,}01$ para C. Além disso, nenhuma correlação significativa foi detectada pelo coeficiente de máxima informação (MIC), cujos valores são $\approx 0{,}07$ para ambas as relações. Essas análises indicam que nossos resultados obtidos com *embedding dimensions* $d_x = d_y = 2$ não são enviesados pelas dimensões das imagens.

rísticas visuais opostas, entre as quais temos o conceito de *linear* versus *pictórica*. Obras de arte "lineares" são compostas por formas claras e bem delineadas, enquanto nas obras "pictóricas," os contornos são sutis e não tão bem definidos, mesclando diferentes partes da imagem e passando a ideia de fluidez. Por sua vez, Riegl considera uma classificação baseada na dicotomia entre o *háptico* e o *óptico*. Segundo Riegl, obras "hápticas" retratam os objetos como entidades discretas, tangíveis, isoladas e circunscritas. Por outro lado, as obras "ópticas" representam objetos inter-relacionados no espaço profundo, explorando luz, cor e efeitos de sombra para criar a ideia de um espaço aberto e contínuo.

Acreditamos que as noções de ordem/simplicidade versus desordem/complexidade nos arranjos dos *pixels* das imagens capturadas pelo plano complexidade-entropia codificam, ao menos em parte, esses conceitos. Imagens formadas por partes distintas e bem delineadas resultam em muitas repetições de poucos padrões ordinais distintos. Consequentemente, obras lineares/hápticas devem ser descritas por valores pequenos para H e valores grandes para C. Por outro lado, imagens compostas por partes inter-relacionadas, delimitadas por bordas suaves e borradas produzem padrões mais aleatórios e, portanto, obras pictóricas/ópticas devem resultar em valores maiores para H e menores para C. É importante mencionarmos que os conceitos duais de Wölfflin e Riegl são formas limitantes de representação, os quais demarcam extremos contendo uma possível escala de todas as possibilidades intermediárias de representação [112]. Nesse sentido, o caráter contínuo dos valores de H e C pode ajudar os historiadores da arte a graduarem essa escala de possibilidades.

Nesse contexto, investigamos se a escala definida pelo valores de H e C é capaz de revelar alguma propriedade dinâmica da arte. Para responder essa questão, estimamos os valores médios de H e C após agrupar as imagens de acordo com a data de produção da obra de arte. Devido ao fato das obras de arte não estarem uniformemente distribuídas no tempo (conforme vimos na figura 2.1A), escolhemos intervalos de tempo contendo aproximadamente o mesmo número de imagens em cada janela. A figura 2.6 mostra a evolução conjunta dos valores de H e C ao longo dos anos, ou seja, as mudanças no plano complexidade-entropia. Essa figura revela uma tendência clara, cuja robustez é evidenciada na figura 2.7. A trajetória dos valores de H e C mostra que as obras de arte produzidas entre os séculos IX e XVII são, em média, mais regulares/ordenadas do que as que foram criadas entre os séculos XIX e a primeira metade do século XX. Além disso, as obras de arte produzidas após 1950 são ainda mais regulares/ordenadas do que as pertencentes aos dois períodos anteriores. Por fim, observamos que o ritmo das mudanças no plano complexidade-entropia se intensifica após o século XIX, um período que coincide com a emergência de vários estilos artísticos, como o Neoclassicismo e o Impressionismo, e também com o aumento na diversidade de constaste de cores reportado por Lee *et.al.* [107].

As três regiões destacadas na figura 2.6 possuem uma grande correspondência com as principais divisões da história da arte. O primeiro período (retângulo preto) corresponde à

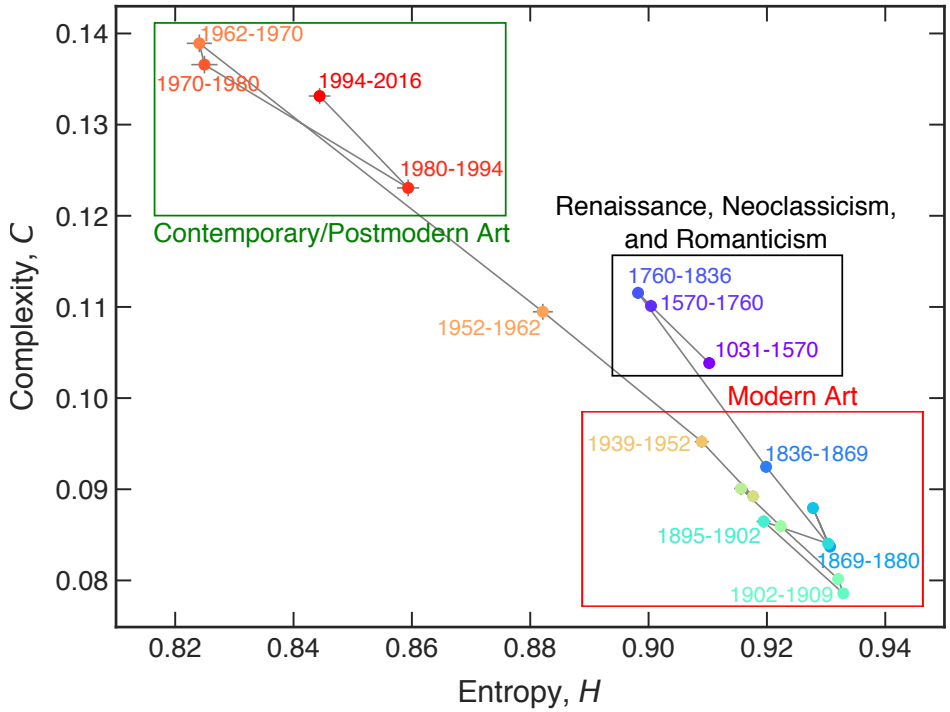

Figura 2.6: Quantificando a evolução das obras de arte ao longo da história da arte. Essa figura mostra a evolução temporal dos valores médios da entropia de permutação H e da complexidade estatística C (plano complexidade-entropia). Cada ponto corresponde aos valores médios de H e C para um dado intervalo temporal (que é mostrado no gráfico). As barras de erro representam o erro padrão da média. As regiões destacadas indicam diferentes períodos da arte (preto: Renascença, Neoclassicismo e Romantismo; vermelho: Arte Moderna; verde: Arte Pós-Moderna/Contemporânea). É interessante notar que o plano complexidade-entropia identifica corretamente os diferentes períodos da arte e as transições entre eles.

Arte Medieval, Renascença, Neoclassicismo e Romantismo, os quais se desenvolveram até por volta de 1850. O segundo período (retângulo vermelho) corresponde à Arte Moderna, marcada pelo surgimento do Impressionismo (por volta de 1870) e pelo desenvolvimento de vários movimentos vanguardistas (como Cubismo, Expressionismo e Surrealismo) durante as primeiras décadas do século XX. Finalmente, o último período corresponde à transição entre a Arte Moderna e a Arte Pós-Moderna/Contemporânea. A data específica que marca o começo do período Pós-Moderno ainda é objeto de grande debate entre os especialistas em arte [113]. No entanto, há algum consenso que a Arte Pós-Moderna surge com o desenvolvimento da Arte Pop nos anos 1960 [113].

Levando a analogia entre o plano complexidade-entropia e os conceitos de Wölfflin e Riegl adiante, a transição entre a arte produzida antes do Modernismo e a Arte Moderna representa uma mudança nos modos de representação de linear/háptico para pictórico/óptico. Esse resultado está de acordo com a ideia de que as obras de arte da Renascença, Neoclassicismo e Romantismo geralmente representam objetos de forma bem distinta e separados por

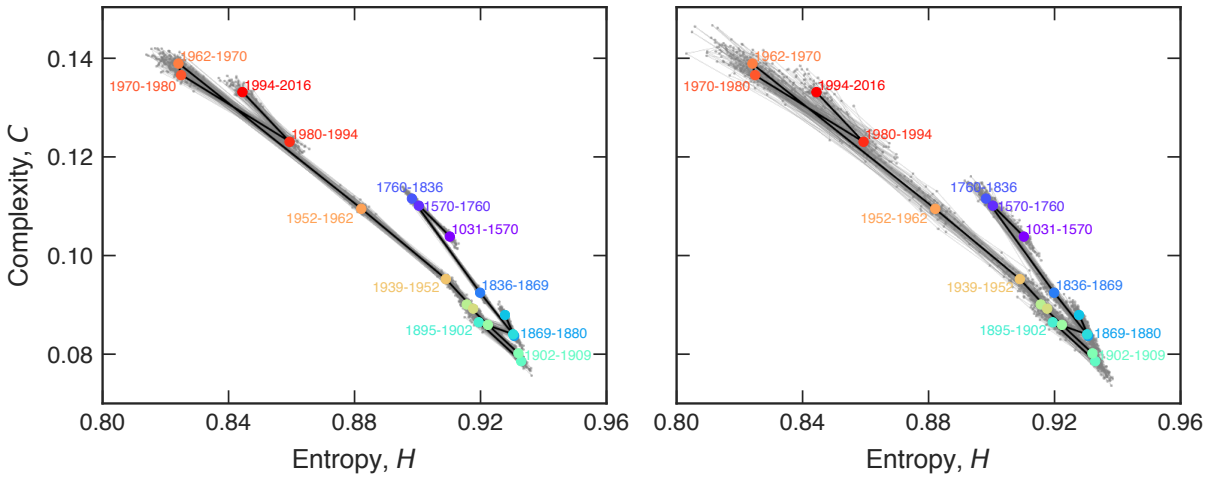

Figura 2.7: Robustez da tendência observada na evolução temporal de H e C ao reamostrar o conjunto de dados. Cada curva cinza corresponde aos valores médios de H e C obtidos a partir de uma amostra aleatória composta de 30% (painel à esquerda) e 10% (painel à direita) das imagens em nosso conjunto de dados. Realizamos o processo de reamostragem por 100 vezes. As curvas pretas mostram a tendência média obtida ao utilizar todo o conjunto de dados, como mostrado na figura 2.6. Notamos que as tendências históricas observadas para os valores médios de H e C são robustas com respeito à reamostragem, sendo que o mesmo padrão é obtido ao utilizar apenas 10% de todas as imagens.

superfícies planas [110, 114, 115]. Por outro lado, estilos Modernos (como Impressionismo, Fauvismo, Pontilhismo e Expressionismo) são marcados pelo uso de pinceladas mais soltas e borradas para evitar a criação de bordas pronunciadas [110, 114, 115]. É interessante notar que a transição entre a Arte Moderna e Pós-Moderna é marcada por uma mudança ainda mais rápida e intensa do pictórico/óptico para o linear/háptico. Esse fato parece concordar com a ideia Pós-Moderna de que a arte deve ser instantaneamente reconhecível, feita de objetos ordinários e marcada pelo uso de bordas largas e bem definidas [114, 115].

Ambas concepções sobre história da arte propostas por Wölfflin e Riegl consideram que o desenvolvimento da arte se dá por meio de uma mudança nos modos de representação de linear/háptico para pictórico/óptico, descrição que concorda com a primeira transição observada na figura 2.6. No entanto, para Riegl [116] esse desenvolvimento ocorre por meio de um processo contínuo e único, enquanto Wölfflin possui uma concepção cíclica dessa transição, a qual parece ser mais consistente com o comportamento dinâmico de H e C. Por outro lado, essa concepção cíclica não é compatível com o comportamento local persistente dessas mudanças no plano complexidade-entropia, ou seja, os valores de H e C não são sempre alternados. Na verdade, estudos mais recentes de historiadores da arte, como o trabalho de Gaiger [112], argumentam que nenhuma dessas concepções é válida ao analisar todo o desenvolvimento da história da arte. Para Gaiger, as categorias duais de Wölfflin e Riegl devem ser tratadas como conceitos puramente descritivos e não ligadas a uma mudança

particular ao longo do tempo.

2.5 Distinguindo estilos artísticos com o plano complexidade-entropia

Uma outra questão interessante diz respeito à capacidade do plano complexidade-entropia para distinguir os diferentes estilos artísticos em nosso conjunto de dados. Para investigar essa possibilidade, calculamos os valores médios de H e C após agrupar as imagens por estilo. Limitamos essa análise aos 92 estilos contendo mais de 100 imagens cada (que totalizam $\approx 90\%$ dos dados) para obtermos valores confiáveis para as médias. Na figura 2.8, observamos que os estilos artísticos estão bem espalhados no plano complexidade-entropia. Os valores médios de H e C são significativamente diferentes para a maioria ($\approx 92\%$) das comparações par a par realizada por meio da estatística do teste-t, como observamos na figura 2.9. Ainda assim, observamos alguns estilos com valores médios estatisticamente indistinguíveis entre si.

Também notamos que a localização dos estilos concorda com a tendência geral observada na evolução temporal dos valores médios de H e C, no sentido que a maioria dos estilos Pós-Modernos estão localizados em uma região de valores menores para a entropia e maiores para a complexidade quando comparados aos estilos Modernos (como Expressionismo, Impressionismo e Fauvismo). Esse arranjo mapeia os diferentes estilos em uma escala contínua, na qual os valores extremos refletem a dicotomia dos modos de representação linear/háptico versus pictórico/óptico. Entre os estilos que possuem os maiores valores para C e os menores para H, temos o Minimalismo, *Hard Edge Painting* e Campo de Cores, os quais são marcados pelo uso de elementos simples e bem delimitados por transições abruptas de cores [114, 115]. Já os estilos que possuem os menores valores para C e os maiores para H (como Impressionismo, Pontilhismo e Fauvismo) são caracterizados pelo uso de pinceladas difusas e borradas, assim como pela mistura de cores para evitar a criação de bordas pronunciadas [114, 115].

2.6 Estrutura hierárquica dos estilos artísticos

Podemos considerar que os valores de H e C capturam o grau de similaridade entre vários estilos artísticos levando em conta a ordem local dos *pixels* das imagens. Utilizando o valor médio dessas medidas de complexidade para cada estilo artístico, investigamos uma possível organização hierárquica dos estilos com respeito a essa ordem local. Para isso, consideramos a distância euclidiana entre um par de estilos no plano complexidade-entropia como uma medida de similaridade entre eles. Assim, quanto menor a distância entre dois estilos artísticos, mais significativa é a similaridade entre eles; por outro lado, pares de

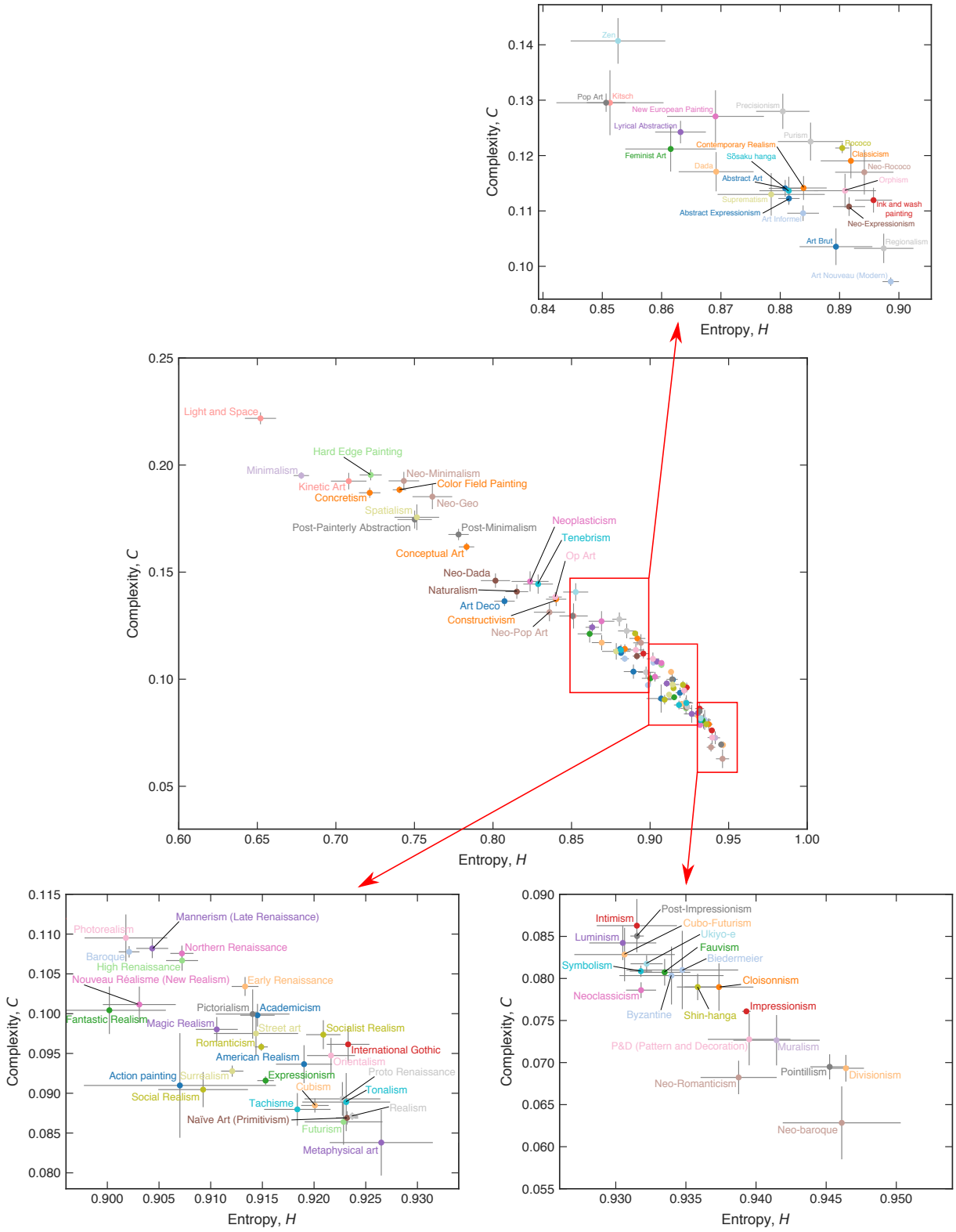

Figura 2.8: Distinguindo entre diferentes estilos artísticos com o plano complexidade-entropia. Os pontos coloridos representam os valores médios de H e C para cada um dos 92 estilos com mais de 100 imagens em nosso conjunto de dados. As barras de erro representam o erro padrão da média.

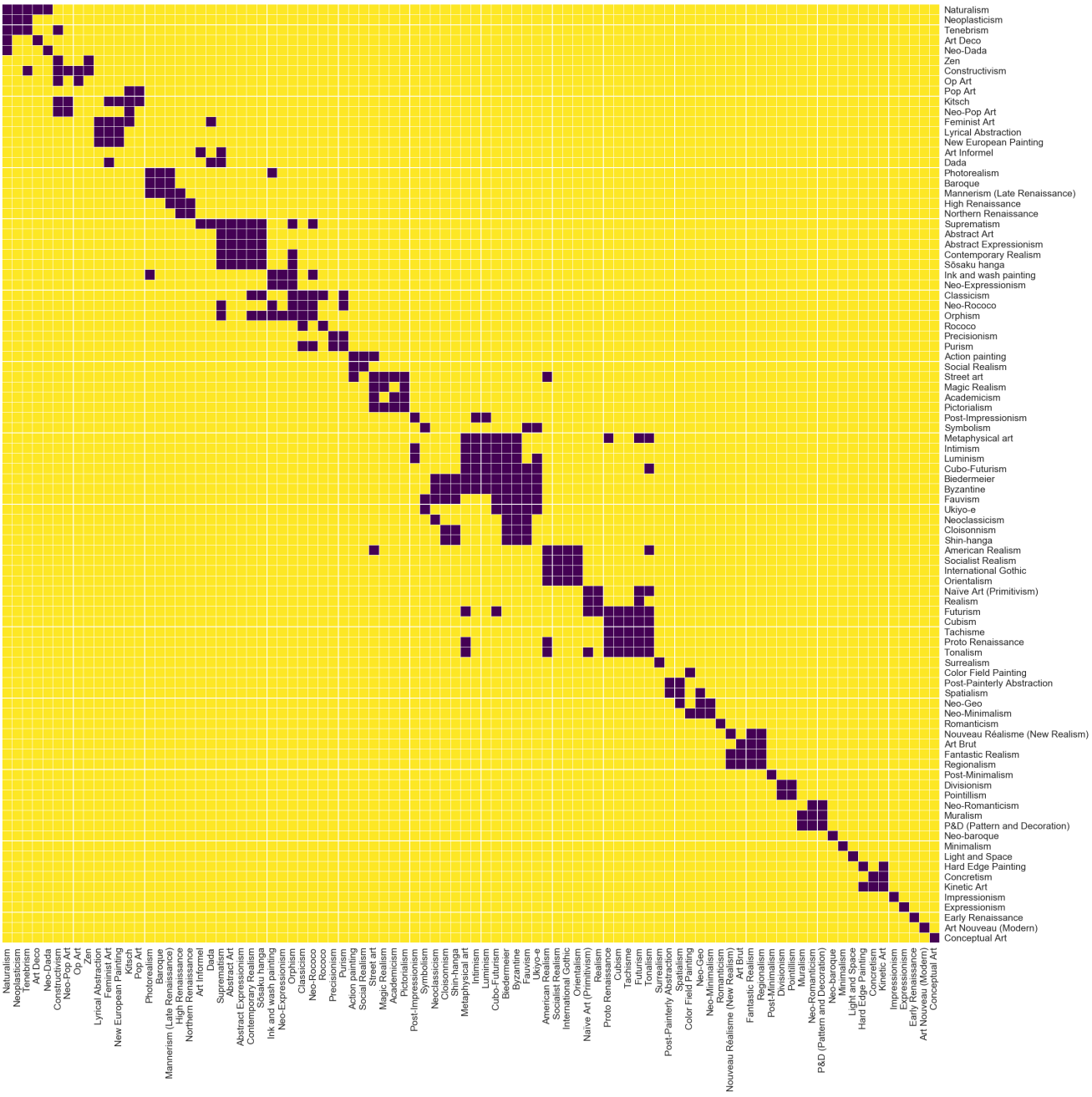

Figura 2.9: Os valores médios de H e C são significativamente diferentes entre a maioria dos estilos. O gráfico da matriz mostra o resultado da estatística do teste-t de duas amostras que compara a diferença para os valores médios de H e C entre todos os possíveis pares de estilos. Consideramos a correção de Bonferroni [117] para levar em conta o fato que estamos realizando múltiplos testes de hipótese. As células amarelas indicam as comparações nas quais a hipótese nula é rejeitada com 95% de confiança (isto é, há uma diferença significativa entre os valores médios de H e/ou C para os dois estilos), enquanto as células roxas indicam comparações nas quais a hipótese nula não pode ser rejeitada (ou seja, nesse caso, não é observada uma diferença significativa entre os valores médios de H e C para os dois estilos). Notamos que a hipótese nula é rejeitada em 91,7% das comparações par a par.

estilos separados por distâncias maiores são considerados mais dissimilares uns dos outros. A figura 2.10A mostra a matriz dessas distâncias, na qual já é possível observar a formação de grupos de estilos.

Para investigar de forma sistemática esses agrupamentos de estilos artísticos, empregamos o método da variância mínima proposto por Ward [118] para construir um dendrograma que representa a matriz das distâncias. Esse método de aprendizagem estatística é um procedimento de agrupamento hierárquico que usa a variância intragrupo como critério para associar pares de objetos. A figura 2.10B mostra esse dendrograma, revelando uma relação intrincada entre os estilos artísticos em nosso conjunto de dados. A distância limiar que utilizamos para segmentar o dendrograma e, assim, determinar o número de grupos, foi obtida maximizando o coeficiente de silhueta [119]. Esse coeficiente quantifica a consistência do procedimento de agrupamento e é definido pelo valor médio de

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{2.1}$$

para todos os i estilos, sendo a_i a distância média intragrupo (também chamada de coesão que, em nosso caso, é o valor médio das distâncias que separam o estilo i dos demais estilos de seu grupo), b_i a distância média do estilo i para os estilos do grupo mais próximo (a chamada separação) e $\max(a_i, b_i)$ representa o maior valor entre essas duas quantidades.

Do modo como é definido, s_i varia entre -1 e 1, sendo que quanto maior for seu valor médio sobre todos os estilos, melhor é a configuração do agrupamento. De maneira intuitiva, para $s_i \approx 1$ é necessário que $a_i \ll b_i$, ou seja, como a_i mede a dissimilaridade média dentro de seu próprio grupo, um valor pequeno indica que o estilo i é bem próximo dos demais de seu grupo e, então, foi bem agrupado. De maneira similar, temos $s_i \approx -1$ se $b_i \ll a_i$, sendo assim, o estilo i está mais próximo do grupo vizinho do que de seu próprio grupo; nesse caso, o estilo i não está bem agrupado.

Assim, uma maneira de encontrar o melhor valor para a distância limiar que fragmenta o dendrograma em grupos é procurar pelo valor que maximiza o coeficiente global de silhueta sobre todos os estilos. Essa análise é mostrada na figura 2.11, na qual temos o valor desse coeficiente em função da distância limiar. Observamos que o coeficiente de silhueta global é máximo (0,57) quando a distância limiar é igual a 0,03. Sendo assim, esse valor é uma escolha natural para segmentar o dendrograma da figura 2.10B. De fato, nessa figura, a linha tracejada indica esse limiar e as diferentes cores do dendrograma mostram os 14 agrupamentos de estilos artísticos obtidos por esse procedimento.

Esses grupos refletem parcialmente a localização temporal dos diferentes estilos artísticos e a evolução mostrada na figura 2.6. Em particular, vários estilos que surgiram juntos ou próximos no tempo são similares com relação ao arranjo local de seus *pixels* e, portanto, pertencem ao mesmo grupo. Por exemplo, os primeiros cinco grupos da figura 2.10B contêm

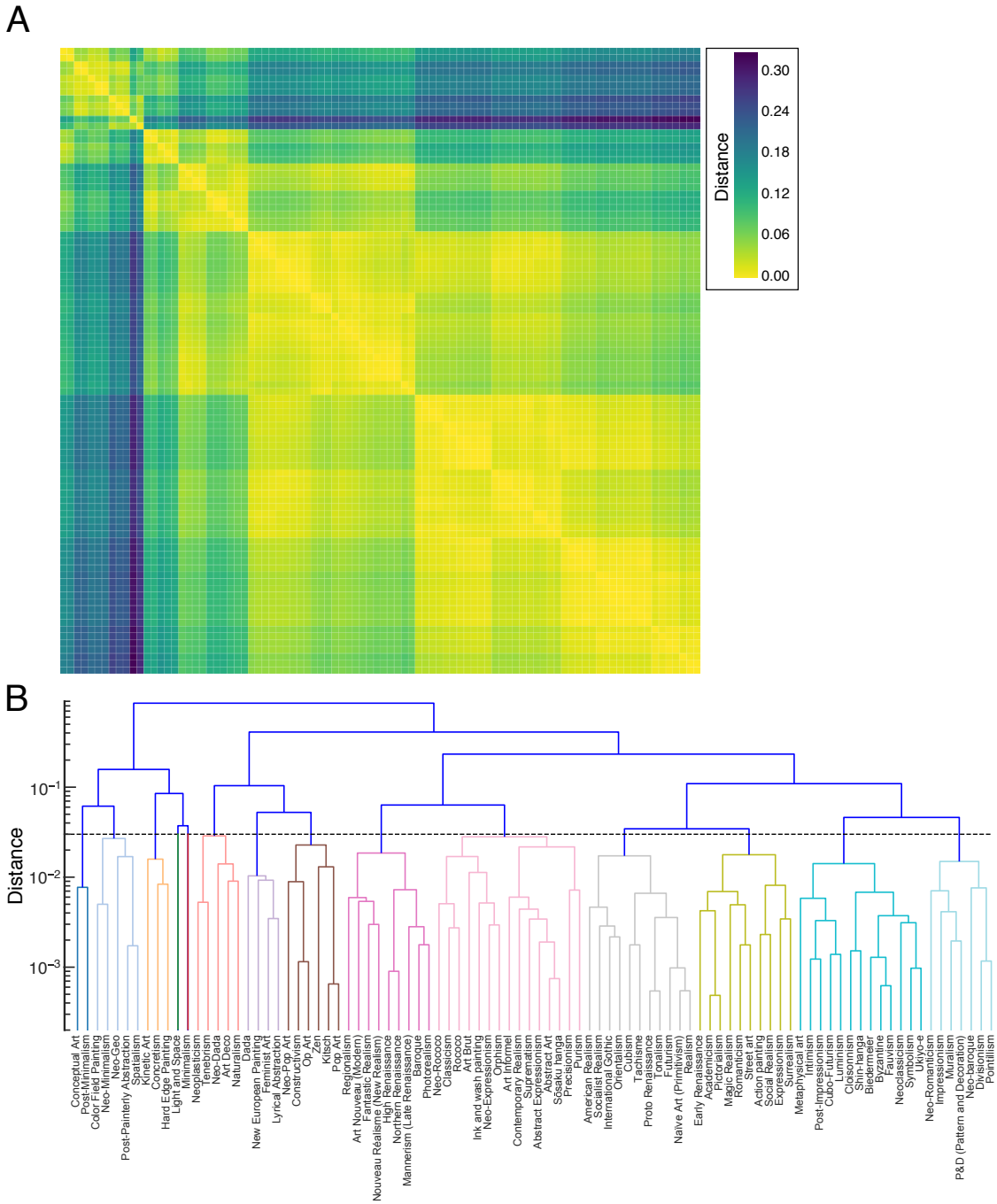

Figura 2.10: Organização hierárquica dos estilos artísticos. (A) Gráfico da matriz das distâncias euclidianas entre cada par de estilos no plano complexidade-entropia. (B) Dendrograma representando a matriz das distâncias obtido aplicando o método da mínima variância proposto por Ward [118]. Os 14 grupos de estilos indicados pelos ramos coloridos foram obtidos segmentando o dendrograma na distância limiar de 0,03. Esse valor maximiza o coeficiente de silhueta [119] e, portanto, define naturalmente o número de grupos em nosso conjunto de dados. A ordem das linhas e colunas no gráfico da matriz das distâncias é a mesma utilizada no dendrograma.

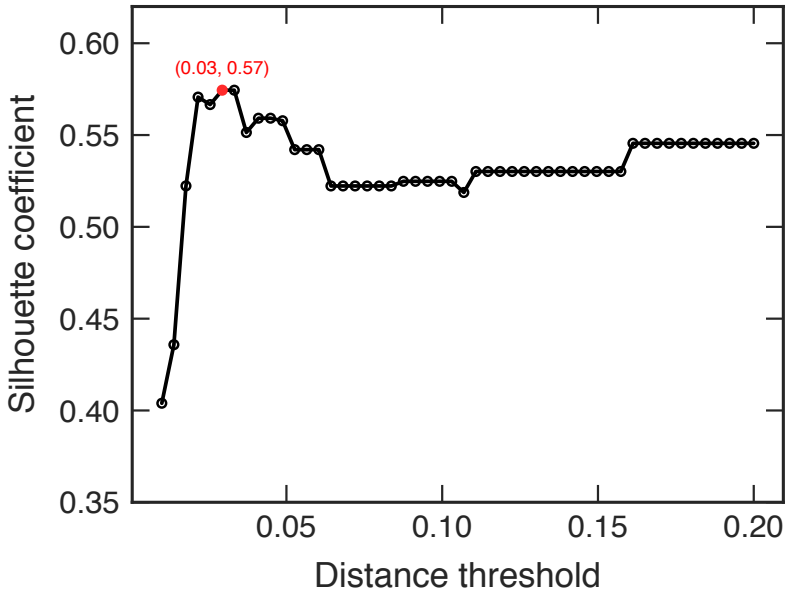

Figura 2.11: Coeficiente de silhueta para os grupos obtidos segmentando o dendrograma da figura 2.10B em diferentes distâncias limiares. Esse coeficiente quantifica a qualidade dos agrupamentos e varia entre -1 e 1. Quanto maior o valor do coeficiente de silhueta, mais consistente são os grupos formados. Portanto, ao encontrar a distância limiar que maximiza o coeficiente de silhueta, estamos maximizando a qualidade dos agrupamentos obtidos a partir do dendrograma. Podemos observar que o coeficiente de silhueta possui um valor máximo (0,57) para a distância limiar de 0,03, valor usado para segmentar o dendrograma e definir o número de grupos da figura 2.10.

principalmente estilos Pós-Modernos. Por outro lado, esses grupos e a estrutura hierárquica correspondente organizam os estilos com respeito aos modos de representação na escala delimitada pela dicotomia linear/háptico versus pictórico/óptico. Esse fato é mais evidente ao examinar os grupos em ambos os extremos de ordem e regularidade no plano complexidade-entropia. O grupo mais a direita da figura 2.10B, por exemplo, contém estilos que usam pinceladas relativamente curtas e evitam a criação de bordas pronunciadas. Essa característica é particularmente evidente em obras do Impressionismo, Pontilhismo e Divisionismo, mas também está presente no Neo-Barroco e Neo-Romantismo, assim como nos trabalhos de muralistas e nas pinturas abstratas do estilo P&D (*Pattern and Decoration*). Um aspecto notório desse último estilo é a pintura de padrões (como em tecidos estampados). O P&D é considerado por muitos uma "reação" ao Minimalismo e a Arte Conceitual (que estão localizados no outro extremo do plano complexidade-entropia), principalmente por evitar composições restritas pelo uso de modulações sutis das cores, como observamos nos trabalhos de Robert Zakanitch (`www.wikiart.org/en/robert-zakanitch`), considerado um dos fundadores do P&D. À medida que nos movemos para grupos caracterizados por valores altos de complexidade e baixos de entropia, observamos o agrupamento de estilos marcados pela presença de bordas pronunciadas e padrões bastante contrastantes, geralmente forma-

dos por partes isoladas distintas ou combinadas com materiais não relacionados. Esse é o caso do grupo contendo Op Arte, Arte Pop e Construtivismo, e também do grupo formado por Arte Cinética, *Hard Edge Painting* e Concretismo [114, 115].

Uma maneira de verificar a significância desses grupos é comparando o agrupamento mostrado na figura 2.10 com uma abordagem baseada nas similaridades entre o conteúdo textual das páginas da Wikipédia de cada estilo. Para isso, obtemos o conteúdo textual dessas páginas e extraímos as 100 principais palavras-chave para cada uma aplicando a estatística conhecida por *term frequency-inverse document frequency* (TF-IDF) [120]. A estatística dessa técnica tem como objetivo refletir o quão importante uma palavra é para um documento em um *corpus* (coleção de documentos). O valor do TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece em um documento e é compensado pelo número de documentos na coleção que contém a palavra, o que pondera o fato de que algumas palavras geralmente aparecem mais frequentemente. Consideramos o inverso de 1 mais o número de palavras-chave compartilhadas entre dois estilos como uma medida de similaridade entre eles. Portanto, estilos que não possuem nenhuma palavra-chave em comum estão a uma "distância" máxima igual a 1, enquanto estilos que compartilham várias palavras estão a uma "distância" menor.

Utilizando um procedimento de agrupamento hierárquico similar ao que foi usado na figura 2.10, obtemos 24 grupos de estilos artísticos na análise dos textos da Wikipédia, os quais são mostrados na figura 2.12. Esse número de grupos é muito maior do que os 14 que foram obtidos no plano complexidade-entropia. No entanto, ambos os agrupamentos compartilham similaridades que podem ser quantificadas utilizando as métricas de avaliação de agrupamentos homogeneidade h, completeza c e medida v. Homogeneidade perfeita ($h = 1$) indica que todos os grupos obtidos a partir dos textos da Wikipédia contêm somente estilos que pertencem aos mesmos grupos obtidos via plano complexidade-entropia. Por outro lado, completeza perfeita ($c = 1$) indica que todos os estilos que pertencem ao mesmo grupo obtido a partir do plano complexidade-entropia estão agrupados no mesmo grupo obtido por meio dos textos da Wikipédia. A medida v é a média harmônica entre h e c, ou seja, $v = 2hc/(h+c)$. Ao calcular essas medidas, encontramos $h = 0{,}49$, $c = 0{,}40$ e $v = 0{,}44$, os quais são valores significativamente maiores do que os que são obtidos a partir de um modelo nulo, no qual o número de palavras-chave compartilhadas é escolhido aleatoriamente a partir de uma distribuição uniforme entre 0 e 100 ($h_{rand} = 0{,}45 \pm 0{,}02$, $c_{rand} = 0{,}35 \pm 0{,}01$ e $v_{rand} = 0{,}38 \pm 0{,}01$; valores médios sobre 100 realizações). Sendo assim, as similaridades entre os agrupamentos obtidos por meio das duas abordagens não podem ser explicadas ao acaso. Esse resultado indica que, apesar do caráter local de nossas medidas de complexidade, os valores de H e C refletem parcialmente o significado de algumas palavras-chave utilizadas para descrever estilos artísticos.

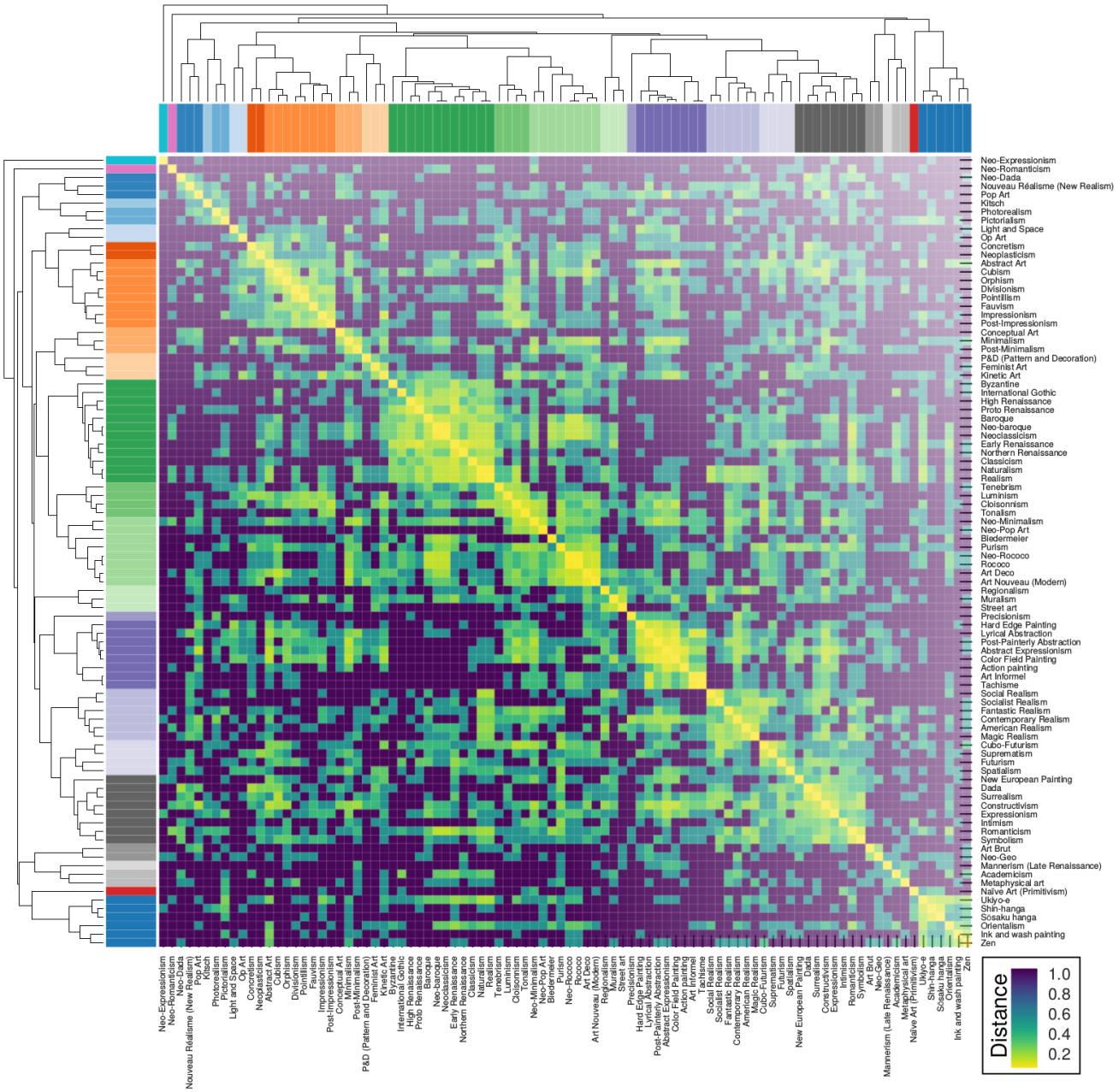

Figura 2.12: Organização hierárquica dos estilos de acordo com as palavras-chave extraídas das páginas da Wikipédia contendo a descrição de cada estilo. Para cada um dos 92 estilos diferentes que possuem ao menos 100 imagens, obtemos o conteúdo textual de suas páginas na Wikipédia. Esses textos foram processados utilizando a abordagem conhecida por *term frequency-inverse document frequency* (TF-IDF) [120] e as 100 principais palavras-chave foram obtidas para cada estilo. Definimos a "distância" entre dois estilos como o inverso de 1 mais o número de palavras compartilhadas entre dois estilos. Essa figura mostra o gráfico da matriz dessas distâncias bem como a representação em dendrograma. Os ramos coloridos indicam os 24 grupos de estilos obtidos ao segmentar o dendrograma na distância limiar que maximiza o coeficiente de silhueta.

2.7 Prevendo estilos artísticos

Outra maneira de quantificar a informação contida nos valores de H e C é tentar prever o estilo de uma imagem baseado somente nesses dois valores. Para isso, implementamos quatro algoritmos de aprendizado de máquina bem conhecidos [71, 121] (k-vizinhos mais próximos, floresta aleatória, máquina de vetores de suporte (SVM) e rede neural) para a tarefa de prever o estilo das imagens dos 20 estilos que possuem mais de 1.500 obras de arte cada. Para cada método, estimamos as curvas de validação para um intervalo de valores dos parâmetros principais dos algoritmos utilizando a estratégia de validação cruzada e estratificada (cada subamostra contém aproximadamente a mesma fração de cada classe observada nos dados) com $n = 10$ camadas. A figura 2.13A mostra as curvas de validação para o algoritmo k-vizinhos mais próximos em função do número de vizinhos k. Note que ocorre *overfit* se o número de vizinhos for menor que ≈ 250. Observamos também que o *score* de validação cruzada satura em $\approx 0{,}18$ se o número de vizinhos for maior que 300 e que não ocorre *underfit* para até pelo menos 500 vizinhos.

Outra questão relevante ao empregar métodos de aprendizagem estatística está relacionada à fração do conjunto de dados que é necessária para treinar apropriadamente o modelo, conforme já discutimos na seção 1.5. Para investigar essa questão, utilizamos novamente uma estratégia de validação cruzada e estratificada com $n = 10$ para estimar as curvas de aprendizagem. A figura 2.13B mostra os *scores* de treino e de validação cruzada para o algoritmo de k-primeiros vizinhos, na qual notamos que ambos aumentam com o tamanho do conjunto de treino. No entanto, o ganho é muito pequeno quando mais de $\approx 50\%$ dos dados são utilizados para treinar o modelo. A figura 2.14 mostra resultados análogos aos apresentados na figura 2.13 para os outros três algoritmos de aprendizado de máquina.

Combinamos a análise prévia com um algoritmo de *grid search* (uma varredura de parâmetros em uma grade predefinida), para estimar a melhor combinação de parâmetros que maximiza o desempenho de cada método de aprendizagem estatística. Na figura 2.13C, observamos que os quatro algoritmos possuem desempenhos similares, todos exibindo precisão próxima de 18%. Além disso, comparamos essas precisões com as obtidas a partir de dois classificadores *dummy*, isto é, previsões feitas ao acaso. No classificador *dummy* estratificado, as previsões dos estilos são geradas ao acaso mas respeitando a fração dos estilos no conjunto de dados. No caso do classificador *dummy* uniforme, as previsões são uniformemente aleatórias. Os resultados da figura 2.13C mostram que todos os algoritmos de aprendizado de máquina possuem uma precisão significativamente maior do que a obtida ao acaso.

Esse resultado, portanto, confirma que os valores de H e C carregam informações importantes sobre o estilo de cada obra de arte. Apesar disso, a precisão alcançada é bem modesta para aplicações práticas e, de fato, existem outras abordagens que são mais precisas. Por exemplo, Zujovic *et al.* [122] obtiveram precisão de $\approx 70\%$ em uma tarefa de classificação

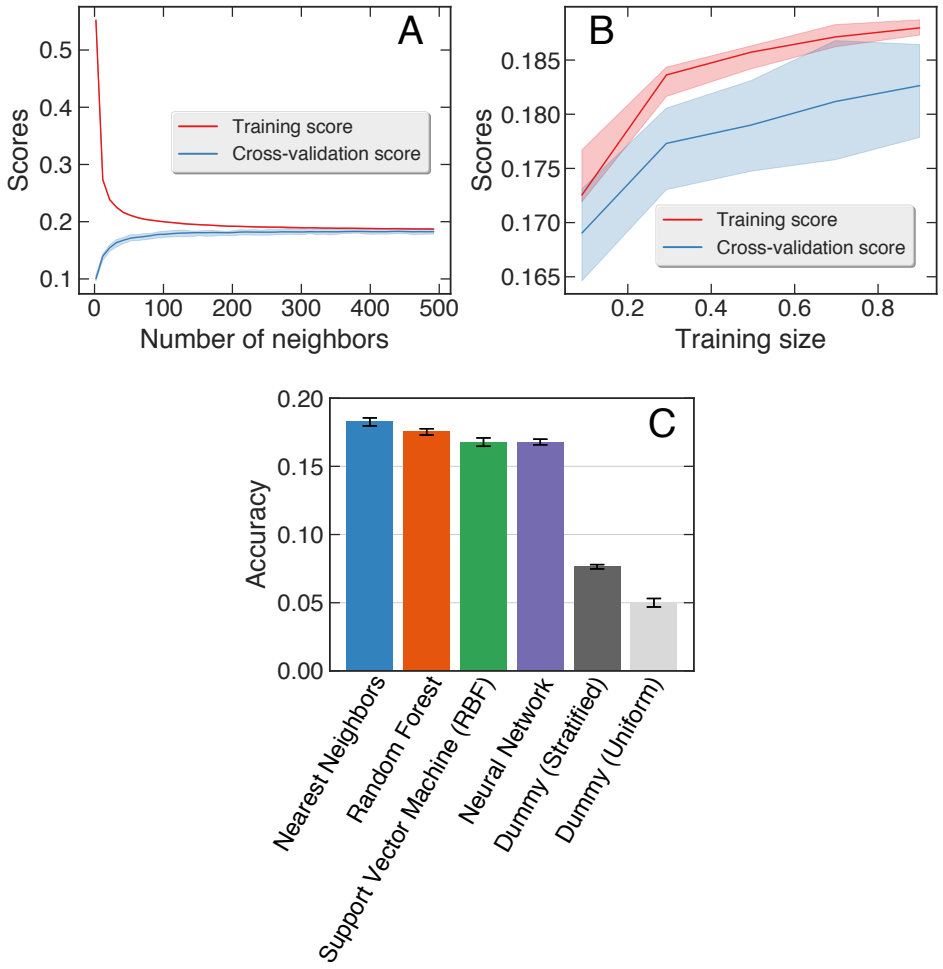

Figura 2.13: Prevendo estilos artísticos com algoritmos de aprendizagem estatística. (A) *Scores* de treino e validação cruzada do algoritmo k-vizinhos mais próximos em função do número de vizinhos k. Notamos que ocorre *overfit* para valores do número de vizinhos menores que 250, mas não notamos aumento na precisão para valores maiores e nem *underfit* para até 500 vizinhos. (B) Curva de aprendizagem, isto é, os *scores* de treino e validação cruzada em função do tamanho do conjunto de treino (fração de todo o conjunto de dados) para o algoritmo de k-vizinhos mais próximos com o número de vizinhos igual a 400. Não observamos melhora significativa no *score* de validação cruzada quando mais de 50% dos dados são utilizados para treinar o modelo. Em ambos os gráficos, as regiões sombreadas indicam os intervalos de confiança de 95% obtidos com uma estratégia de validação cruzada com $n = 10$ camadas. (C) Comparação entre 4 algoritmos de aprendizagem estatística diferentes (k-vizinhos mais próximos, floresta aleatória, máquinas de vetores de suporte e rede neural) e também a precisão obtida a partir de dois classificadores *dummy* (estratificado: gera previsões aleatórias respeitando a fração dos estilos no conjunto de dados; uniforme: as previsões são uniformemente aleatórias). As barras de erro representam o erro padrão da média. Os quatro classificadores possuem precisão similar ($\approx 18\%$) e todos superam significativamente os classificadores *dummy*. Esses resultados são baseados nos 20 estilos com mais de 1.500 imagens cada, embora resultados similares são obtidos quando incluímos os outros estilos.

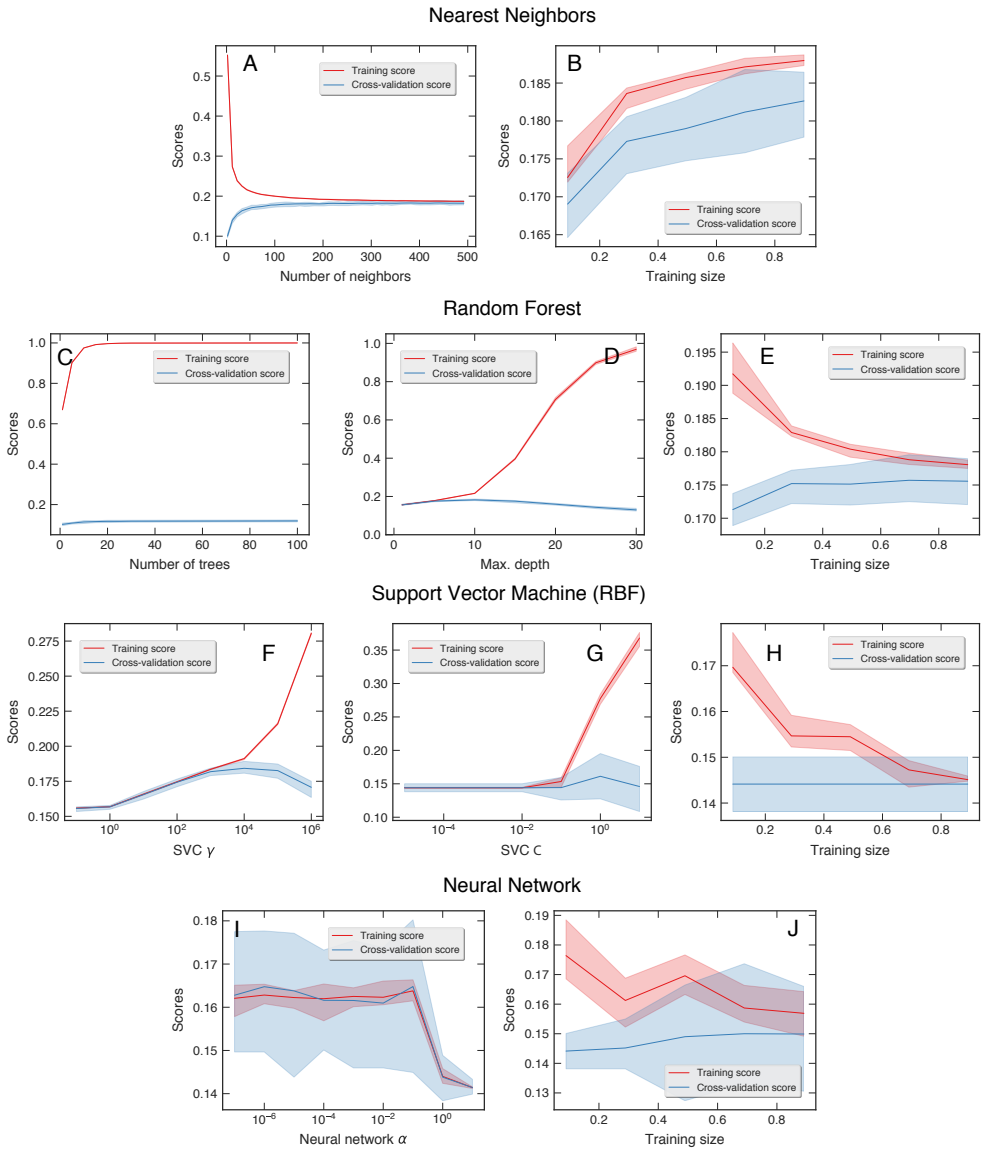

Figura 2.14: *Scores* de treino e validação cruzada obtidos a partir dos quatro algoritmos de aprendizado de máquina que foram utilizados para prever os estilos em função dos parâmetros principais e do tamanho do conjunto de treino. Os painéis (A) e (B) mostram os resultados para o algoritmo k-vizinhos mais próximos. Os painéis (C), (D) e (E) mostram os *scores* para o algoritmo floresta aleatória. Os parâmetros principais, nesse caso, são o número de árvores na floresta e a profundidade máxima dessas árvores. Os painéis (F), (G) e (H) mostram os *scores* para o classificador de vetores de suporte (SVC) com um *kernel* de base radial (RBF). O parâmetro γ é associado à largura do *kernel* RBF e C' é o parâmetro de penalidade. Os painéis (I) e (J) mostram os resultados para o algoritmo de rede neural (um modelo *perceptron* com uma única camada oculta). O parâmetro α é a chamada penalidade L2 e o número de neurônios é igual a 100. As precisões médias reportadas na figura 2.13C foram obtidas para $k = 400$ vizinhos; $\gamma = 10^4$ e $C' = 0{,}1$; número de árvores $= 400$ e a profundidade máxima $= 5$; e $\alpha = 10^{-4}$. Todos os algoritmos estão implementados na biblioteca *scikit-learn* e as curvas de aprendizagem foram estimadas utilizando os melhores parâmetros de cada modelo.

com 353 pinturas de 5 estilos e Argawal *et al.* [123] reportaram uma precisão de $\approx 60\%$ em uma tarefa de classificação com 3.000 pinturas de 10 estilos. No entanto, nossos resultados não podem ser diretamente comparados com os desses trabalhos, já que esses utilizam um conjunto de dados bem menor, com poucos estilos e também várias características das imagens, enquanto nossas previsões são baseadas somente em duas características. Nesse sentido, nossa abordagem representa uma severa redução de dimensionalidade, pois imagens com aproximadamente 1 milhão de *pixels* são representadas por apenas dois números relacionados à ordem local dos *pixels* das imagens. Nesse contexto, uma precisão de 18% em uma tarefa de classificação com 20 estilos e mais de 100.000 obras de arte não pode ser negligenciada. Além do mais, a natureza local de H e C fazem essas medidas de complexidade bastante rápidas, fáceis de serem paralelizadas e escaláveis do ponto de vista computacional. Assim, além de mostrar que o plano complexidade-entropia contém informações importantes sobre os estilos artísticos, acreditamos que os valores de H e C, combinados com outras características das imagens, podem conduzir a uma melhor precisão nessa tarefa de classificação desafiadora.

2.8 Conclusão

Nesse capítulo, apresentamos uma caracterização em grande escala de um conjunto de dados composto de aproximadamente 140 mil imagens digitais de obras de arte produzidas ao longo do último milênio da história da arte. Nossas análises são baseadas em duas medidas de complexidade relativamente simples (entropia de permutação H e complexidade estatística C) e diretamente relacionadas aos padrões ordinais dos *pixels* dessas imagens. Essas medidas mapeiam o grau de ordem local dessas obras em uma escala de ordem/desordem e simplicidade/complexidade que reflete, ao menos em parte, as descrições qualitativas de obras de arte propostas por Wölfflin e Riegl. Em particular, argumentamos que os limites dessa escala correspondem aos dois extremos dos modos de representação propostos por esses historiadores da arte, isto é, a dicotomia entre linear/háptico ($H \approx 0$ e $C \approx 0$) e pictórico/óptico ($H \approx 1$ e $C \approx 0$).

Investigando o comportamento dinâmico dos valores médios das medidas de complexidade empregadas, encontramos uma trajetória evolutiva clara e robusta para a arte no plano complexidade-entropia. Essa trajetória é caracterizada por transições que concordam com as principais divisões da história da arte. Essas transições podem ser classificadas como sendo de linear/háptico para pictórico/óptico (antes e depois da Arte Moderna) e de pictórico/óptico para linear/háptico (a transição entre a Arte Moderna e Pós-Moderna), mostrando que cada um desses períodos históricos apresenta um grau distinto de entropia e complexidade. Por um lado, as concepções sobre história da arte de Wölfflin em termos de uma transição cíclica entre linear e pictórico não concordam com a persistência temporal dos valores de H e C, nem

com o escrutínio crítico de Gaiger [112] e de outros historiadores da arte contemporâneos. Por outro lado, elas são consistentes com a evolução global mostrada no plano complexidade-entropia. Para Wölfflin, a transição de linear para pictórico é governada por uma "lei natural no mesmo sentido que o crescimento físico" e "determinar essa lei seria um problema central, o problema central da história da arte" [110]. No entanto, o retorno para o linear "certamente reside em circunstâncias externas" [110] e, no contexto da figura 2.6, não é difícil imaginar que a transição do período Moderno para o Pós-Moderno tem relação com o fim da Segunda Guerra Mundial, o evento que geralmente marca o início do Pós-Modernismo nos livros de história.

Além de revelar esse aspecto dinâmico da arte, os valores de H e C são capazes de distinguir entre diferentes estilos artísticos de acordo com o grau médio de entropia/complexidade das obras de arte. Enfatizamos que a localização de cada estilo no plano complexidade-entropia reflete parcialmente a dualidade linear/háptico versus pictórico/óptico e, dessa maneira, pode ser considerada como uma régua para quantificar o uso desses modos de representação opostos. Além do mais, as distâncias entre pares de estilos no plano complexidade-entropia representam uma medida de similaridade com relação a esses conceitos da história da arte. Utilizando essas distâncias, encontramos que estilos artísticos diferentes podem ser hierarquicamente organizados e agrupados de acordo com as suas posições no plano. Verificamos que esses grupos refletem bem o conteúdo textual das páginas da Wikipédia utilizadas para descrever cada estilo. Esses grupos também refletem algumas similaridades entre os estilos, principalmente com relação à presença de transições suaves/difusas ou bem definidas/abruptas. Quantificamos a informação contida nessas medidas de complexidade por meio de uma tarefa de classificação na qual o estilo da imagem é predito utilizando apenas os valores de H e C. A taxa de precisão dessa tarefa de aprendizagem estatística é de aproximadamente 18%, valor que supera a precisão obtida por classificadores *dummy* e confirma que essas duas medidas carregam informações significativas sobre o estilo das obras.

O fato dessas medidas de complexidade serem estritamente baseadas em uma escala espacial local das obras de arte, naturalmente faz com que não sejam capazes de capturar toda a pluralidade e complexidade da arte. No entanto, nossos resultados demonstram que medidas simples inspiradas em Física podem ser conectadas a conceitos propostos por historiadores da arte e, mais importante ainda, que essas medidas realmente carregam informações relevantes sobre obras de arte, seus estilos e sua evolução. No contexto da metáfora de Wölfflin sobre a evolução da arte descrita nas suas palavras como: "Uma inspeção mais próxima, em breve, certamente mostrará que a arte mesmo hoje jamais retornou ao ponto em que um dia esteve, mas que somente um movimento em espiral concordaria com os fatos..." [110], podemos considerar o plano complexidade-entropia como uma das possíveis projeções da espiral de Wölfflin.

CAPÍTULO 3

Estimando propriedades físicas a partir de imagens de texturas de cristais líquidos

Técnicas baseadas em imagens são ferramentas essenciais para investigar uma série de propriedades em diferentes materiais. Cristais líquidos estão entre esses materiais que são frequentemente investigados por meio de métodos ópticos e de processamento de imagens. Apesar disso, pouca atenção tem sido voltada ao problema de extrair propriedades físicas de cristais líquidos diretamente de imagens das texturas desses materiais. Neste capítulo, buscamos reduzir essa lacuna por meio de duas abordagens [33, 34]. A primeira combina as medidas de entropia e complexidade estatística de permutação com métodos de aprendizagem de máquina. Já a segunda, emprega redes convolucionais neurais, um método de aprendizagem profunda (*deep learning*) muito utilizado em tarefas que envolvem imagens e que possui a vantagem de dispensar a extração manual de características das imagens.

3.1 Introdução

Técnicas baseadas em imagens são ferramentas importantes e amplamente utilizadas para investigar várias propriedades de diversos materiais [124]. Essas técnicas usualmente são não-destrutivas e particularmente convenientes para lidar com materiais biológicos e outros materiais complexos [125]. Cristais líquidos estão entre esses materiais frequentemente estudados por meio de métodos ópticos e de processamento de imagens [126]. Isso ocorre porque cristais líquidos são materiais birrefringentes e, sendo assim, um microscópio óptico polarizado já acessa algumas de suas propriedades importantes, incluindo birrefringência e espessura da amostra [127].

Apesar do uso extensivo de abordagens baseadas em imagens ópticas no estudo de cristais líquidos, pouca atenção tem sido voltada ao problema de extrair parâmetros físicos diretamente das imagens desses materiais. Essa questão é importante dado que vários parâmetros físicos de cristais líquidos são obtidos somente por meio do ajuste de modelos teóricos a resultados experimentais, uma tarefa usualmente complicada e que pode demandar muito tempo. Exemplos desses parâmetros incluem o parâmetro de ordem microscópico, do qual vários outros parâmetros que caracterizam a fase nemática são dependentes [126], e o comprimento do passo em cristais líquidos colestéricos. Esse último, pode ser facilmente obtido por um microscópio óptico quando o eixo helicoidal fica perpendicular à direção de visualização [128], mas não pode ser estimado a partir de arranjos experimentais mais comuns, nos quais o eixo helicoidal está orientado paralelamente em relação à direção de visualização [129].

Nesse contexto, a caracterização de cristais líquidos baseada em imagens pode se beneficiar bastante das diversas técnicas de aprendizagem estatística [71]. Essas abordagens estão disponíveis desde os anos 1990, mas foi somente durante a última década que esses métodos ganharam popularidade impressionante em várias áreas da ciência, principalmente naquelas em que revelar padrões significativos presentes nos dados é tarefa fundamental. Naturalmente, a Física não é exceção e, de fato, existem vários trabalhos recentes que empregam algoritmos de aprendizado de máquina para estudar sistemas físicos [130–134]. O uso de métodos de aprendizagem profunda (*deep learning*) em Física [135–137], Química [138–140] e Ciência dos Materiais [141–145] é ainda mais recente. O desenvolvimento e a disseminação dos métodos de aprendizado de máquina (sejam do tipo *shallow* ou *deep learning*) combinados com a disponibilidade crescente de grandes conjuntos de dados têm sido reconhecidos como o "quarto paradigma da ciência" [146] e também como a "quarta revolução industrial" [147], tendo grande potencial para valorizar ainda mais o papel dos métodos computacionais em pesquisas aplicadas e fundamentais.

A recência no uso de métodos de aprendizagem estatística também faz com que muitas áreas tenham tirado pouca ou nenhuma vantagem dessas abordagens. Esse também é o caso das pesquisas sobre cristais líquidos [126] que, a despeito da tradição no uso de técnicas baseadas em imagens, têm explorado muito pouco abordagens de aprendizagem estatística. Nesse capítulo, contribuímos para reduzir essa lacuna com duas investigações voltadas para determinar propriedades de cristais líquidos a partir de imagens (texturas) desses materiais.

Tal qual no capítulo anterior, nossa primeira abordagem (seção 3.2) é baseada na avaliação de duas medidas de complexidade (entropia H e complexidade estatística de permutação C) relacionadas ao arranjo dos *pixels* nas texturas, combinadas com algoritmos simples de aprendizagem de máquina para tarefas de classificação e regressão. Demonstramos o potencial dessa abordagem em uma série de aplicações em texturas experimentais e numericamente geradas, a partir das quais parâmetros físicos de cristais líquidos são previstos com alta precisão. Na segunda abordagem (seção 3.3), utilizamos de métodos de *deep learning* – redes

convolucionais neurais profundas – para prever propriedades de amostras de cristais líquidos diretamente de suas texturas ópticas. Otimizando arquiteturas de redes convolucionais simples, verificamos que essa abordagem também é muito eficiente para prever propriedades de cristais líquidos, superando o primeiro procedimento baseado nas medidas de complexidade em algumas tarefas.

3.2 Prevendo propriedades físicas de cristais líquidos com medidas de complexidade estatística

Nesta seção, apresentamos uma abordagem que combina duas medidas de complexidade (entropia e complexidade estatística de permutação) com métodos de aprendizagem estatística, para extrair propriedades físicas de cristais líquidos nemáticos e colestéricos diretamente das imagens de suas texturas. Demonstramos a utilidade e a precisão desse método em uma série de aplicações envolvendo texturas experimentais e simuladas, nas quais propriedades físicas desses materiais (parâmetro de ordem médio, temperatura da amostra e comprimento do passo colestérico) são previstas com precisão bastante significativa.

3.2.1 Prevendo o parâmetro de ordem de texturas nemáticas simuladas

Iniciamos nossa investigação analisando texturas nemáticas geradas por meio de simulações de Monte Carlo do modelo descrito no Apêndice A.1. Exemplos dessas texturas são mostrados na figura 3.1 para diferentes temperaturas reduzidas T_r, isto é, a razão entre a temperatura T e a temperatura crítica $T_c = 1{,}1075$, na qual ocorre a transição da fase nemática ($T < T_c$) para a fase isotrópica ($T > T_c$). Cada textura possui um parâmetro de ordem médio p diferente que depende da temperatura T_r (veja a equação A.3 para detalhes) e nosso objetivo é prever o valor de p diretamente a partir dessas imagens utilizando apenas os valores de H e C.

É interessante notar que essas texturas exibem padrões visualmente distintos entre a fase nemática e isotrópica. Em particular, para $T > T_c$ observamos a emergência de domínios isotrópicos que predominam na textura a medida que a temperatura aumenta. Embora as texturas nemáticas e isotrópicas sejam facilmente distinguíveis umas das outras, mesmo o olho bem treinado de um físico experimental tem dificuldade para tentar distinguir entre texturas nemáticas com temperaturas diferentes (por exemplo, entre $T_r = 0{,}2$ e $T_r = 0{,}3$), assim como entre texturas isotrópicas em temperaturas diferentes.

Apesar da similaridade visual entre as texturas, conjecturamos que os valores de H e C podem distinguir entre essas imagens bem como identificar a transição nemática-isotrópica. Para verificar essa possibilidade, criamos um conjunto de dados composto de várias realiza-

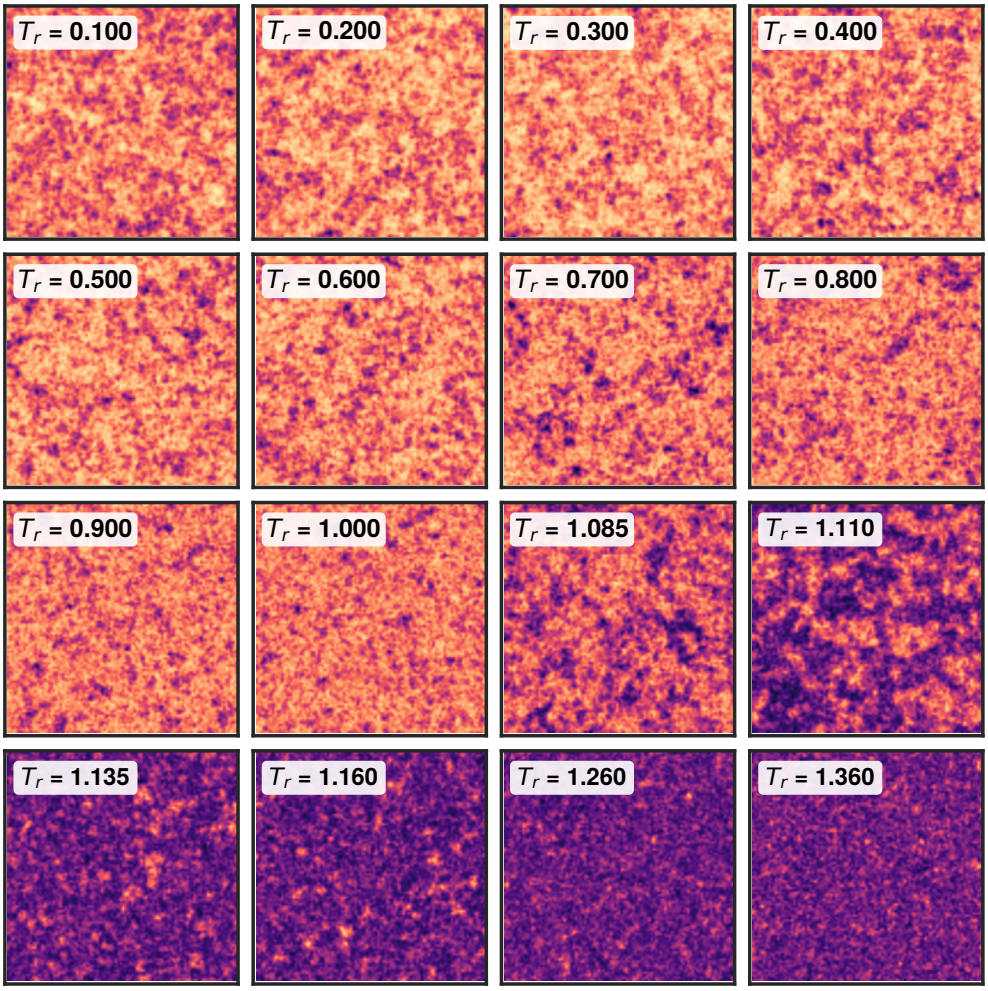

Figura 3.1: Exemplos de texturas de um cristal líquido nemático em diferentes temperaturas e fases. Essas texturas são geradas utilizando o modelo baseado no método de Monte Carlo descrito no Apêndice A.1. Cada textura corresponde a uma temperatura reduzida T_r diferente, $T_c = 1{,}1075$ é a temperatura crítica de transição de fase nemática-isotrópica, como indicado nas imagens. Notamos que texturas de fases diferentes podem ser facilmente distinguíveis, enquanto texturas nemáticas ($T_r < T_c$) em diferentes temperaturas reduzidas são muito similares entre si, assim como as da fase isotrópica ($T_r > T_c$).

ções de texturas nemáticas simuladas para diferente temperaturas T_r e calculamos os valores de H e C para cada uma utilizando as *embedding dimensions* $d_x = d_y = 2$. As figuras 3.2A e 3.2B mostram a dependência dos valores médios de H e C em função da temperatura reduzida T_r. Observamos que H possui uma tendência geral de crescimento com o aumento da temperatura, mas também apresenta um mínimo para $T_r = T_c$. De maneira similar, os valores de C tendem a decrescer com o aumento da temperatura e apresentam um máximo para $T_r = T_c$. A figura 3.2C mostra a dependência do parâmetro de ordem p em função da temperatura reduzida T_r, na qual a temperatura crítica $T_c = 1{,}1075$ é definida como o valor que maximiza a derivada de p com respeito a T_r.

A dependência bem definida dos valores médios de H e C com a temperatura T_r, com-

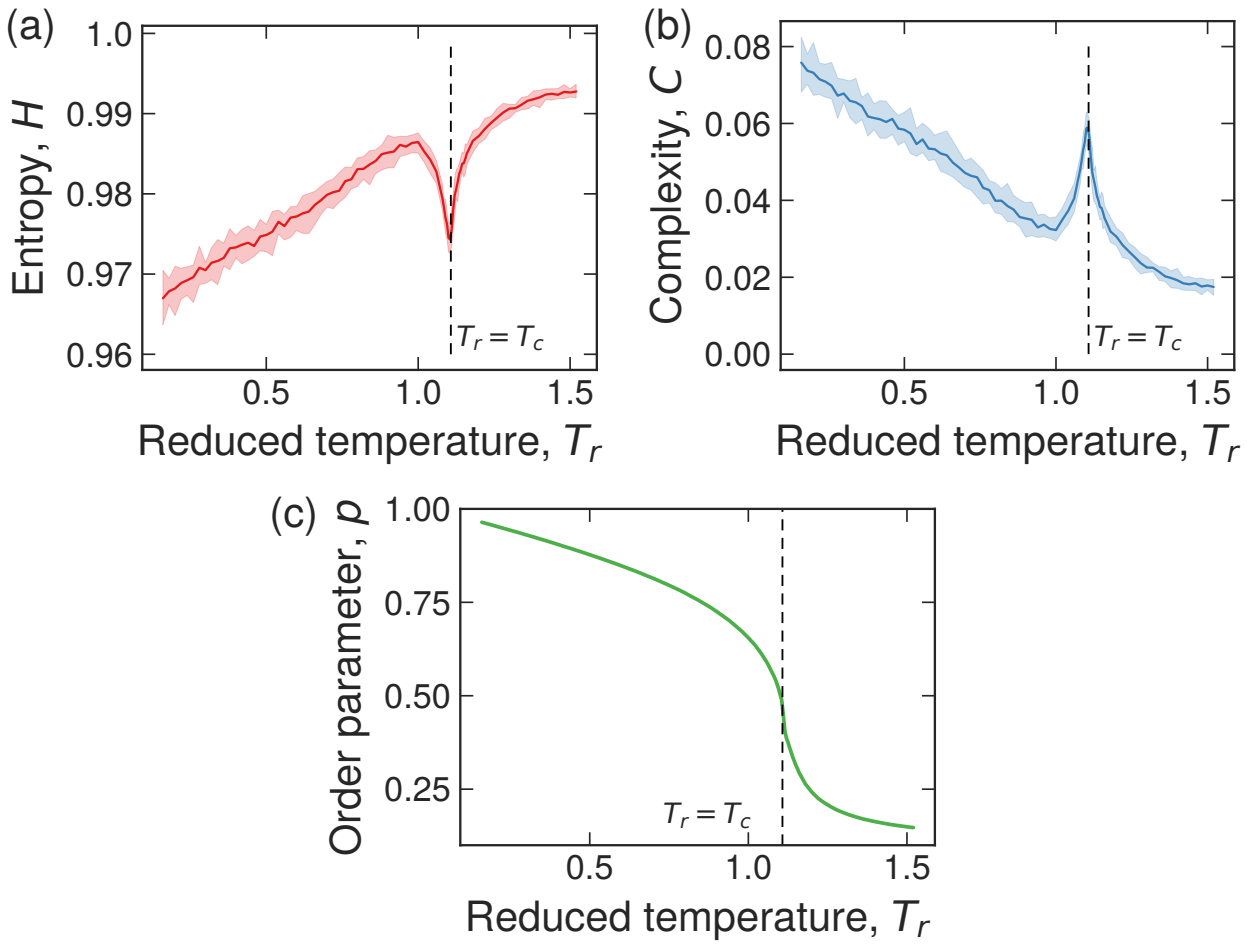

Figura 3.2: Dependência dos quantificadores das imagens e do parâmetro de ordem com a temperatura. (A) Valores da entropia H e (B) da complexidade estatística C de permutação em função da temperatura reduzida T_r. As curvas sólidas representam os valores médios após 50 realizações e as áreas sombreadas são os intervalos de confiança de 95%. (C) Dependência do parâmetro de ordem p em função da temperatura reduzida T_r. Em todos os painéis, a linha tracejada vertical indica a temperatura crítica ($T_c = 1{,}1075$) de transição entre a fase nemática e a isotrópica. Notamos que a transição de fase é claramente e apropriadamente identificada pelos valores extremos das medidas de complexidade.

binada com o fato de que o parâmetro de ordem p também é uma função de T_r, indica que podemos prever os valores de p diretamente a partir das imagens. É interessante notar que os valores de H e C não são unicamente definidos para uma dada temperatura, apresentando uma flutuação aleatória associada ao processo que gera essas texturas. Desse modo, para testar o poder preditivo desses quantificadores das imagens em uma situação mais prática, treinamos o algoritmo de k-vizinhos mais próximos para a tarefa de prever o parâmetro de ordem p baseado nos valores de H e C. Além disso, utilizamos uma variável *dummy* que é 0 no intervalo $T_c - 0{,}05T_c \leq T \leq T_c$, -1 quando $T < T_c - 0{,}05T_c$ e 1 para $T > T_c$. Essa variável *dummy* é necessária devido à relação não biunívoca entre os quantificadores das imagens e a temperatura T_r. Entretanto, vale notar que essa variável *dummy* é completamente obtida

a partir dos valores de H e C, de modo que essa informação também é extraída a partir das imagens.

A figura 3.3A mostra as curvas de validação, ou seja, os *scores* de treino e de validação cruzada em função do número de vizinhos k. Nas tarefas de regressão, os *scores* representam o coeficiente de determinação R^2, e, quanto maior o seu valor, mais explicativo é o modelo. Notamos que ocorre *overfit* para $k < 10$, isto é, o algoritmo é muito complexo e modela até o ruído aleatório do conjunto de treino. Por outro lado, notamos a ocorrência de *underfit* para $k > 20$, o que indica que o método de aprendizagem estatística não é complexo o suficiente para capturar a estrutura subjacente aos dados. A figura 3.3B mostra as curvas de aprendizagem, ou seja, os *scores* em função da fração do conjunto de dados utilizada para treinar o algoritmo (com $k = 15$). Não observamos melhora significativa no *score* de validação cruzada quando mais de 35% dos dados são usados como conjunto de treino. Portanto, o algoritmo de k-vizinhos mais próximos com $k = 15$ atinge uma precisão de $\approx 99\%$ na tarefa de regressão de prever o parâmetro de ordem p baseado somente nos valores de H e C. A figura 3.3C mostra uma comparação entre os valores reais do parâmetro de ordem p em função da temperatura reduzida T_r para uma simulação em particular e os valores previstos pelo algoritmo de k-vizinhos mais próximos com $k = 15$. Notamos que as previsões são muito próximas dos valores reais do parâmetro de ordem, o que confirma a eficiência e a utilidade de nossa abordagem.

3.2.2 Prevendo a temperatura de amostras experimentais

Em outra aplicação, estudamos texturas nemáticas obtidas a partir de amostras do cristal líquido E7, um material comumente empregado na indústria para a produção de *displays*. Esse cristal líquido exibe a transição nemática-isotrópica em $T_c \approx 58$ °C [127]. Detalhes sobre o procedimento experimental são fornecidos no Apêndice A.2. Entretanto, o experimento consiste basicamente em utilizar um microscópio óptico polarizado para tirar fotografias dessas texturas em diferentes temperaturas T. A figura 3.4A mostra exemplos dessas texturas obtidas a partir de uma amostra com a temperatura variando entre 40 °C e 60 °C. Observamos que não existem mudanças visuais significativas nos padrões dessas texturas até a temperatura de 55 °C. Quando a temperatura da amostra excede a temperatura crítica $T_c \approx 58$°C, observamos o surgimento de domínios isotrópicos, o que torna mais fácil a distinção entre essas texturas.

Coletamos dados de seis amostras de cristal líquido E7 seguindo o mesmo protocolo experimental. As figuras 3.4B e 3.4C mostram o comportamento médio da entropia H e da complexidade C em função da temperatura T. Esses resultados são similares àqueles obtidos via simulação de Monte Carlo (veja a figura 3.2), ou seja, H tende a crescer com o aumento de T e possui um mínimo na temperatura crítica. Por outro lado, C possui uma tendência decrescente com o aumento de T e um ponto de máximo na temperatura crítica. Novamente,

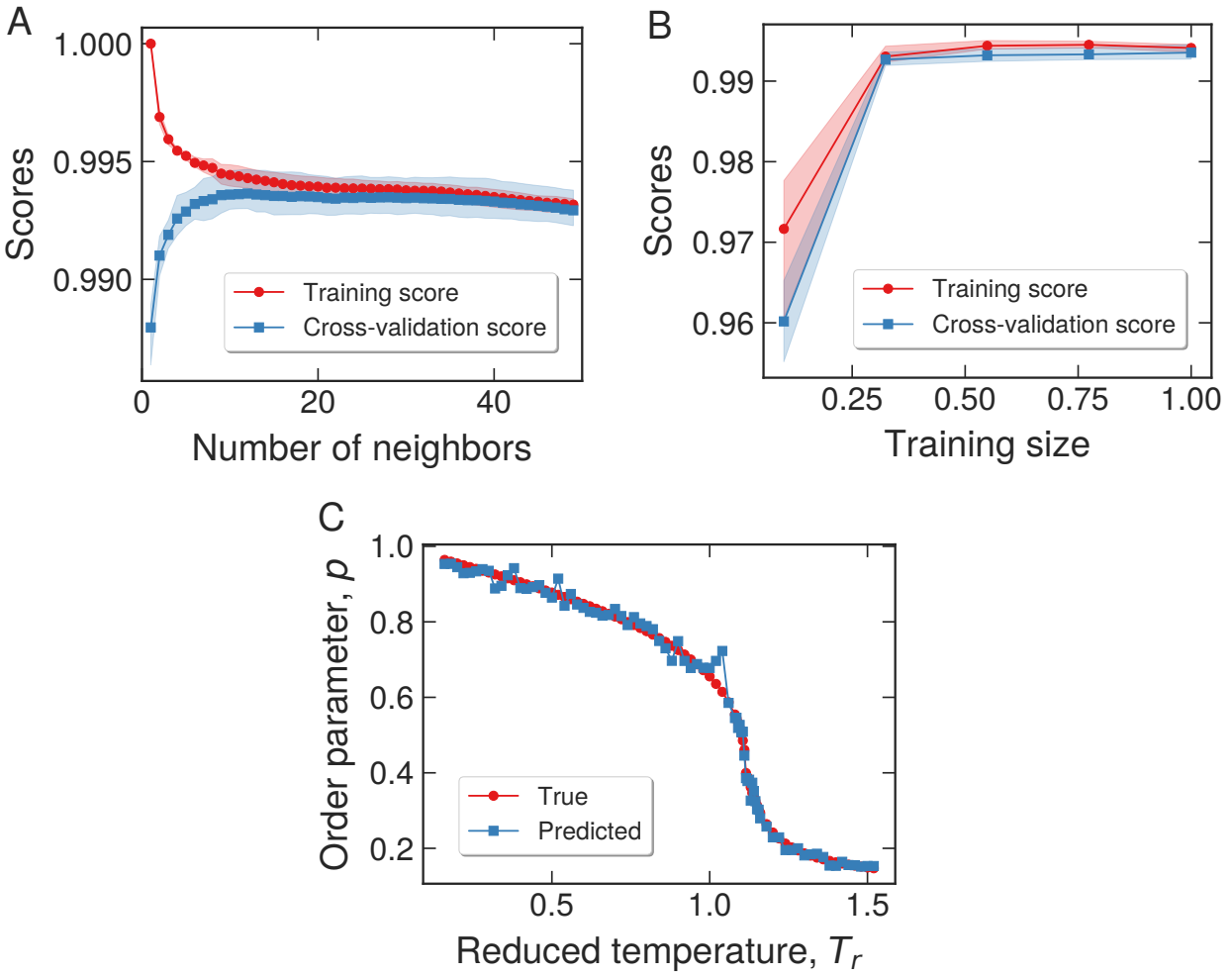

Figura 3.3: Prevendo o parâmetro de ordem p utilizando um algoritmo de aprendizagem estatística. (A) *Scores* de treino e validação cruzada do algoritmo k-vizinhos mais próximos em função do número de vizinhos k. Notamos que ocorre *overfit* para $k < 10$, enquanto para $k > 20$ começa a ocorrer *underfit*. (B) Curvas de aprendizagem, ou seja, os *scores* de treino e de validação cruzada em função do tamanho do conjunto de treino (fração de todo o conjunto de dados usada para treinar o modelo) com $k = 15$. Não observamos melhora significativa no *score* de validação cruzada quando mais de 35% dos dados são usados. As áreas sombreadas em ambos os gráficos representam o intervalo de confiança de 95% para os *scores* obtidos a partir de uma estratégia de validação cruzada com $n = 3$ camadas. Notamos a alta precisão alcançada pelo algoritmo ($\approx 99\%$). (C) Parâmetro de ordem p verdadeiro (círculos) e previsto (quadrados) em função da temperatura reduzida T_r. Essas previsões foram geradas expondo o regressor treinado (com $k = 15$) a um conjunto de texturas nunca apresentado ao algoritmo.

observamos que os valores de H e C são funções bem definidas da temperatura T e capazes de identificar precisamente a transição nemática-isotrópica.

Diferentemente dos resultados anteriores para as texturas simuladas, agora vamos prever a temperatura T da amostra (ao invés do parâmetro de ordem p) diretamente a partir das imagens das texturas experimentais. No entanto, é interessante mencionarmos que com esse

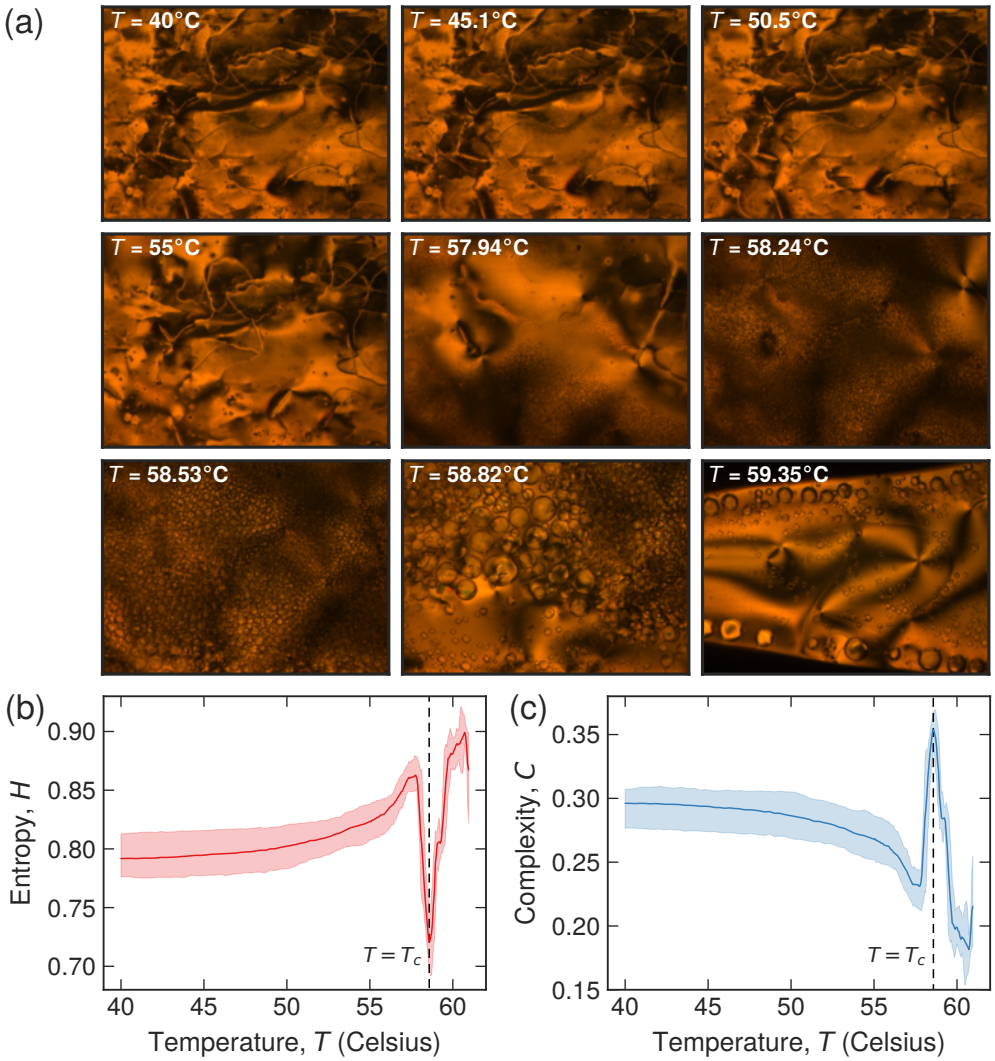

Figura 3.4: Exemplos de texturas do cristal líquido E7 em diferentes temperaturas e fases e a dependência das medidas de complexidade com a temperatura. (A) Texturas experimentais de uma amostra de cristal líquido E7 em diferentes temperaturas. Observamos que praticamente não há mudanças visuais até a temperatura crítica de $T_c \approx 58$ °C. (B) Dependência da entropia H e (C) da complexidade estatística de permutação C com a temperatura. As curvas sólidas representam os valores médios dessas quantidades calculadas para seis amostras. As áreas sombreadas representam o intervalo de confiança de 95%. As linhas verticais tracejadas indicam a temperatura crítica T_c. Notamos que a transição de fase é apropriadamente identificada pelos valores extremos das medidas de complexidade.

arranjo experimental, o parâmetro de ordem pode ser estimado a partir da temperatura [148], de tal forma que prever a temperatura é comparável a prever o parâmetro de ordem.

Para realizar essas previsões, procedemos como no caso simulado, ou seja, treinamos um algoritmo de k-vizinhos mais próximos para a tarefa de regressão de prever os valores de T a partir dos quantificadores das imagens (H e C) e uma variável *dummy* que, como no caso anterior, é 0 no intervalo $T_c - 0{,}05T_c \leq T \leq T_c$, -1 quando $T < T_c - 0{,}05T_c$ e 1 para

$T > T_c$. A figura 3.5A mostra as curvas de validação, na qual observamos a ocorrência de *overfit* para $k = 1$, ao passo que começa a ocorrer *underfit* para $k > 3$. A figura 3.5B mostra as curvas de aprendizagem para $k = 2$, na qual não notamos melhora significativa no *score* de validação cruzada quando o conjunto de treino excede 80% de todo o conjunto de dados. Portanto, o algoritmo de k-vizinhos mais próximos alcança uma precisão de $\approx 93\%$ com $k = 2$. A figura 3.5C ilustra a precisão dessas previsões mostrando um gráfico de dispersão entre os valores previstos para a temperatura e os valores reais. Notamos que essa relação é muito próxima de uma reta 1:1 (indicada pela linha tracejada), o que reforça a qualidade geral das previsões do algoritmo k-vizinhos mais próximos e demonstra o potencial da nossa abordagem em dados experimentais.

3.2.3 Prevendo o comprimento do passo de texturas colestéricas simuladas

Como última aplicação dessa abordagem baseada na entropia e complexidade de permutação, investigamos texturas simuladas de cristais líquidos colestéricos. Esses materiais exibem uma estrutura helicoidal composta de camadas entre as quais o eixo preferencial do diretor varia periodicamente com um período (ou seja, a distância necessária para ocorrer uma rotação completa do eixo diretor) conhecido como passo η. Entre outras propriedades, o passo de um cristal líquido colestérico define o comprimento de onda da luz refletida como consequência da refração de Bragg [126]. Além disso, o comprimento do passo modifica bastante a textura desses materiais. Por exemplo, uma textura colestérica pode imitar uma textura nemática para valores grandes do passo η.

Nosso objetivo, nesse caso, é identificar o comprimento do passo de um cristal líquido colestérico baseado nos valores de H e C obtidos a partir das texturas. Para isso, criamos um conjunto de dados de texturas composto de 100 imagens para cada valor do passo $\eta \in (15, 17, 19, 21, 23, 25, 27, 29, 40)$ nm, com $\eta = 40$ nm sendo grande o suficiente para imitar uma textura nemática. Essas texturas são numericamente obtidas resolvendo o modelo baseado na teoria de Landau-de Gennes [149] descrito no Apêndice A.3. Em particular, foi utilizado um processo de resfriamento rápido (*quenching*) de uma amostra colestérica a partir do estado isotrópico para gerar esse conjunto de texturas. A figura 3.6 mostra os valores de H e C estimados para cinco texturas selecionadas aleatoriamente para cada valor do passo. As inserções nessa figura indicam três texturas típicas para $\eta = 17$ nm, $\eta = 27$ nm e $\eta = 40$ nm. Observamos que embora exista uma superposição, texturas com passos diferentes tendem a ocupar regiões distintas no plano complexidade-entropia, indicando que os valores de H e C são capazes de distinguir entre diferentes texturas colestéricas. Também notamos que quanto maior o valor de η, maior a complexidade e menor a entropia. Portanto, valores grandes para o comprimento do passo produzem texturas mais ordenadas localmente, enquanto valores

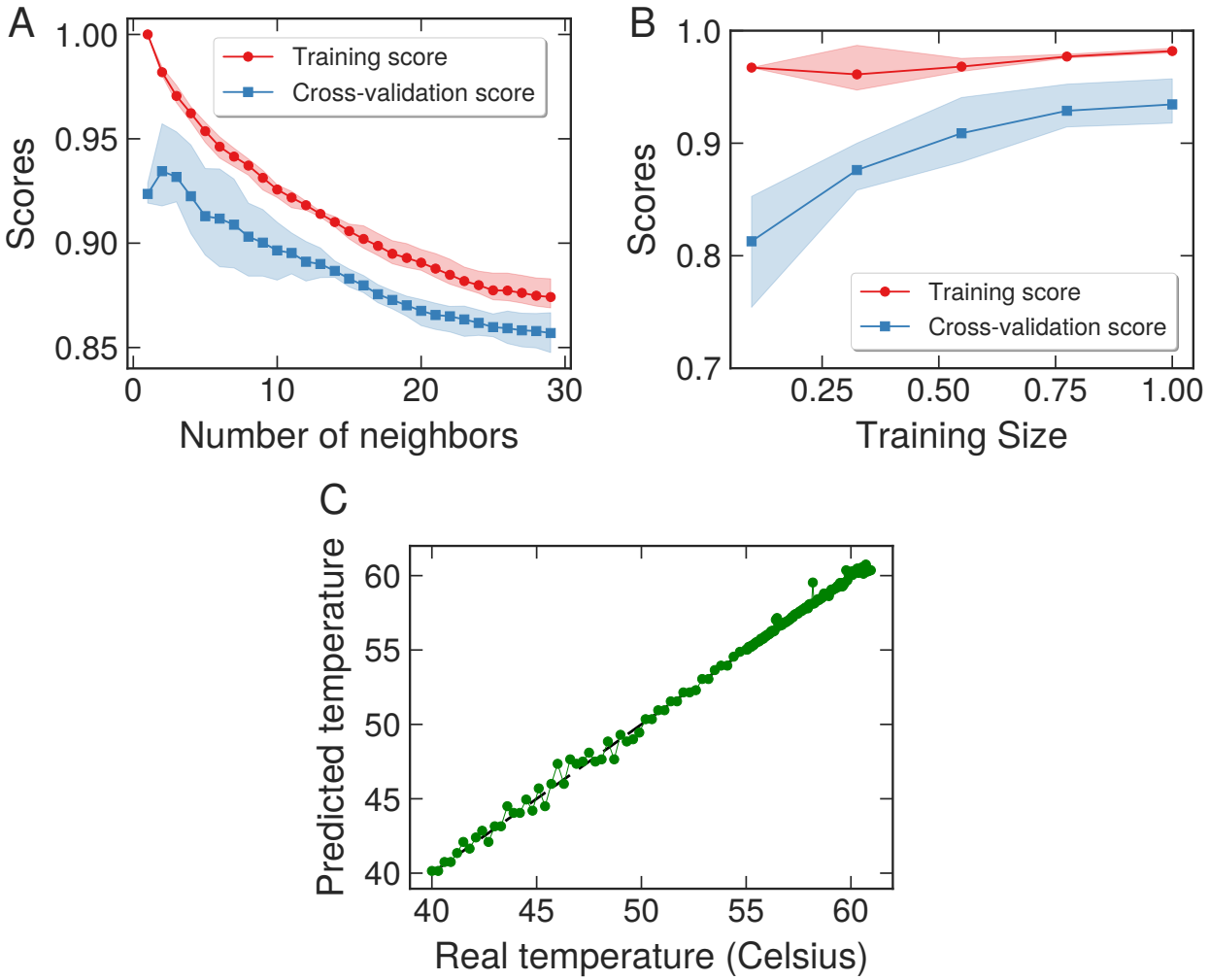

Figura 3.5: Prevendo a temperatura utilizando um algoritmo de aprendizagem estatística. (A) *Scores* de treino e validação cruzada do algoritmo k-vizinhos mais próximos em função do número de vizinhos k. Notamos a ocorrência de *overfit* para $k = 1$, enquanto para $k > 3$ começa a ocorrer *underfit*. (B) Curvas de aprendizagem, ou seja, os *scores* de treino e de validação cruzada em função do tamanho do conjunto de treino (fração de todo o conjunto de dados) com $k = 2$. Não observamos melhora significativa no *score* de validação cruzada quando mais de 80% dos dados são usados para treinar o modelo. As áreas sombreadas em ambos os gráficos representam intervalos de confiança de 95% obtidos a partir de uma estratégia de validação cruzada com $n = 3$ camadas. (C) Temperatura verdadeira e prevista obtidas ao expor o regressor treinado a um conjunto de texturas nunca apresentadas ao algoritmo. A linha tracejada representa a reta de coeficiente angular 1 e intercepto nulo. Observamos o excelente acordo entre as temperaturas verdadeiras e previstas, o que reforça a ótima precisão alcançada pelo método ($\approx 93\%$).

pequenos geram texturas que são localmente mais irregulares.

Treinamos o algoritmo de k-vizinhos mais próximos para a tarefa de classificação de prever os passos com base apenas nos valores de H e C. A figura 3.7A mostra as curvas de validação para esse processo de treinamento. Observamos a ocorrência de *overfit* para o número de vizinhos menor que 15 e, para valores maiores que 20, começa a ocorrer *underfit*.

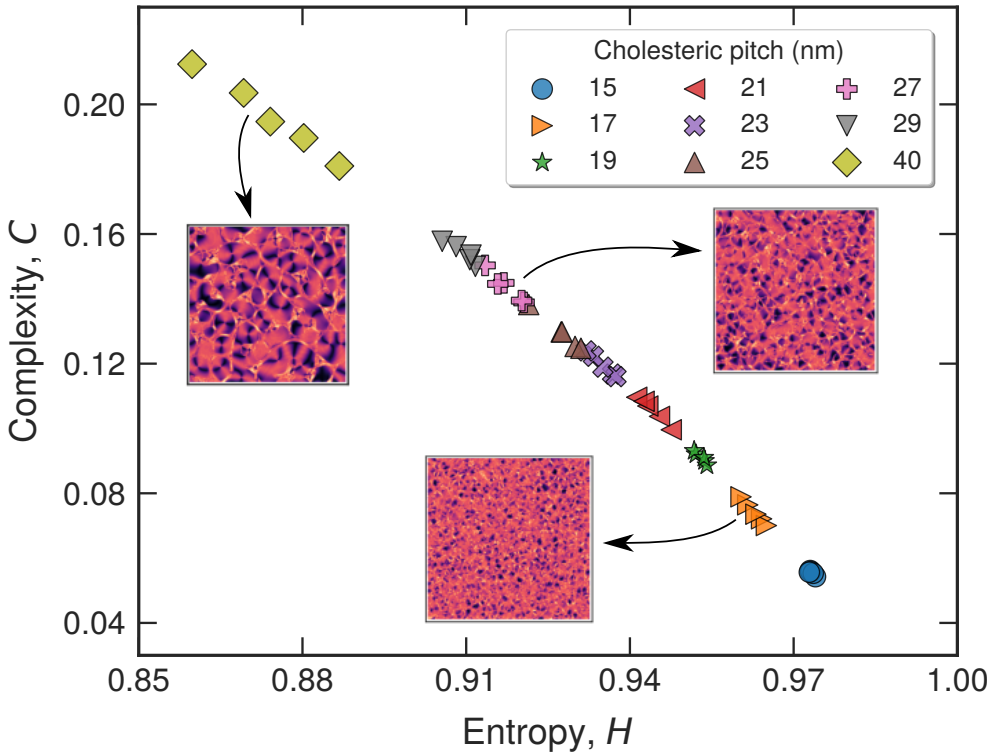

Figura 3.6: Discriminando entre texturas colestéricas com diferentes comprimentos de passos via plano complexidade-entropia. Cada símbolo colorido representa os valores de H e C para 5 realizações de texturas colestéricas (escolhidas ao acaso) com valores de passo diferentes (como indicado pelos símbolos diferentes). Notamos que os valores de H e C se agrupam ao redor de regiões pequenas do plano complexidade-entropia para cada passo colestérico. Também observamos que as texturas com passos curtos estão localizadas em uma região de alta entropia, enquanto aquelas com passos longos estão em uma região de baixa entropia. Esse resultado indica que as texturas com passos maiores são mais ordenadas do que aquelas obtidas com passos curtos (como ilustrado pelas inserções).

Destacamos que esse algoritmo simples atinge precisão de $\approx 85\%$ com $k = 20$. Além disso, as curvas de aprendizagem mostradas na figura 3.7B indicam que $\approx 60\%$ dos dados são suficientes para ajustar o algoritmo aos dados das texturas colestéricas. Também estimamos a matriz de confusão, como mostrado na figura 3.7C. Os elementos f_{ij} dessa matriz representam a fração de texturas com passo η_i que o algoritmo prevê ter passo η_j. Desse modo, um classificador perfeito é representado pela matriz diagonal ($f_{ij} = \delta_{ij}$). De maneira prática, quanto mais próximo de 1 forem os elementos da diagonal de f_{ij}, melhor é o desempenho do classificador. Em nosso caso, observamos que praticamente todos os elementos que não são iguais a zero estão dentro de uma banda diagonal de largura unitária dessa matriz, com a diagonal principal concentrando pelo menos $\approx 72\%$ das previsões. Assim, mesmo quando o algoritmo classifica incorretamente o passo de uma textura (o que ocorre em aproximadamente 15% das previsões), ele tende a prever um valor de passo que é próximo do valor

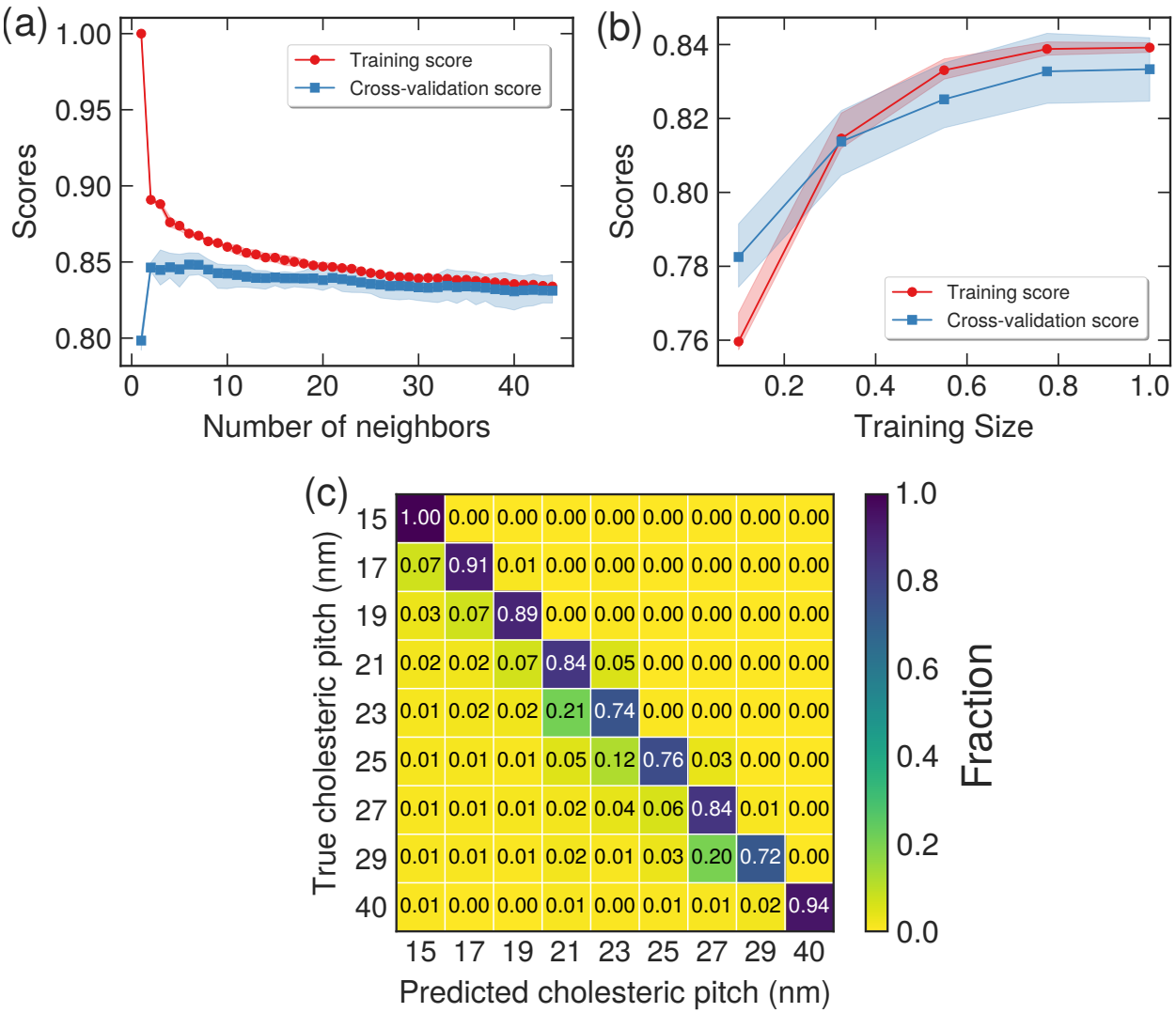

Figura 3.7: Prevendo o comprimento do passo colestérico utilizando um algoritmo de aprendizagem estatística. *Scores* de treino e validação cruzada do algoritmo k-vizinhos mais próximos em função do número de vizinhos k (painel A) e em função do tamanho do conjunto de treino (painel B). As áreas sombreadas em ambos os gráficos representam o intervalo de confiança de 95% obtidos a partir de uma estratégia de validação cruzada com $n = 5$ camadas e com o número de vizinhos igual a 20. Notamos a ocorrência de *overfit* para $k < 15$, enquanto para $k > 20$ começa a ocorrer *underfit*. Também observamos que não há melhora significativa no *score* de validação cruzada quando mais de 60% dos dados são usados para treinar o modelo. (C) Comprimento do passo previsto e verdadeiro. Essa matriz de confusão mostra o bom desempenho alcançado pelo algoritmo, representado pelos valores altos das frações de previsões corretas na diagonal. Em alguns casos, o algoritmo subestima o passo, o que pode ser explicado pela superposição observada no plano complexidade-entropia da figura 3.6.

real, com maior probabilidade de subestimar o valor (note que os elementos que precedem a diagonal principal são maiores do que os que aparecem depois). Esses resultados, portanto, corroboram a utilidade de nossa abordagem para investigar texturas mais complexas.

3.3 Estimando propriedades físicas de cristais líquidos via redes convolucionais neurais

Nessa seção, apresentamos uma outra abordagem em que utilizamos redes convolucionais neurais para estimar propriedades de cristais líquidos diretamente de imagens das suas texturas. A principal diferença dessa abordagem para a anterior está no fato de que ao empregar redes convolucionais neurais não há a necessidade de realizar procedimentos manuais para extrair variáveis preditoras, como foi feito ao calcularmos a entropia e a complexidade de permutação. Ao usar redes convolucionais neurais, os dados de entrada são imagens de texturas de cristais líquidos e os dados de saída são propriedades desses materiais (fase, parâmetro de ordem médio, comprimento do passo e temperatura da amostra). O conjunto de dados utilizado nessas aplicações é o mesmo da seção anterior e consiste em imagens de texturas obtidas a partir de simulações de amostras de cristais líquidos nemáticos e colestéricos, bem como texturas experimentais obtidas de amostras do cristal líquido E7. Como mencionado anteriormente, os Apêndices A.1, A.2 e A.3 fornecem detalhes sobre os procedimentos envolvidos na construção desse conjunto de dados.

Há uma infinidade de possibilidades para as escolhas de arquiteturas de redes convolucionais neurais, as quais envolvem número de camadas, número e tamanho dos filtros e *strides*. Essas escolhas são principalmente empíricas, dependem do tipo dos dados de entrada e muitas vezes são inspiradas por outras arquiteturas que provaram ser bem-sucedidas em tarefas específicas. Contudo, alguns padrões de *design* são comuns em várias arquiteturas de redes convolucionais [150]. Esses padrões incluem parcimônia, simetria, construção incremental de características e estratégias de subamostragem a medida que a rede fica mais profunda [150]. Nossas escolhas específicas ao definir a arquitetura das redes convolucionais usadas nessa seção foram guiadas por esses princípios e em procedimentos de tentativa e erro, além de validação cruzada com alguns parâmetros da rede.

3.3.1 Prevendo a fase de texturas nemáticas simuladas

Em uma primeira aplicação, utilizamos uma rede convolucional neural para detectar se um cristal líquido está na fase nemática ou na fase isotrópica. As texturas simuladas são as mesmas usadas na seção 3.2.1 e apresentam uma transição da fase nemática para a isotrópica em uma temperatura crítica T_c (figura 4.1). Texturas com temperatura abaixo da T_c são entendidas como nemáticas e aquelas com temperatura acima da T_c são consideradas isotrópicas. Essa tarefa de classificação é visualmente simples quando as texturas obtidas estão distantes da temperatura crítica T_c, mas torna-se um desafio quando as texturas estão próximas da temperatura crítica T_c, como já mostramos na figura 3.1. A figura 3.8A ilustra a arquitetura da rede neural que foi utilizada nessa tarefa. Nessa rede, as imagens de entrada

(100 × 100 pixels) passam por dois blocos de convolução 2×2 e camadas de agrupamento (*max pooling*) 2×2, seguidos por duas camadas completamente conectadas (com 32 e 16 nós, respectivamente) e uma camada de saída. Utilizamos a função de ativação *rectified linear unit* (ReLU), que é definida como $f(x) = \max(0, x)$, em todas as camadas convolucionais e nas camadas completamente conectadas, enquanto a camada de saída usa a função de ativação sigmoide, que corresponde à regressão logística.

Separamos 15% do conjunto de dados para avaliação final do modelo (conjunto de teste) e usamos o restante como conjuntos de validação (20%) e de treino (80%). Os parâmetros da rede são otimizados usando o algoritmo de gradiente estocástico Adam [151] (com atualização da taxa de aprendizagem igual a 0,001) e a função de perda é a entropia cruzada binária (comumente usada em tarefas de classificação binária). Para evitar *overfitting*, aplicamos um procedimento de regularização conhecido por *early stopping*, que encerra o processo de treinamento quando a função de perda avaliada no conjunto de validação para de melhorar (ou seja, o erro para de diminuir) durante um intervalo de dez épocas (valor do chamado parâmetro paciência). Também incluímos um termo de penalização na função de perda proporcional a soma dos quadrados dos parâmetros da camada (conhecido como regularização L2, com hiper-parâmetro $\lambda = 0{,}005$, que é a constante de proporcionalidade) sobre todas as camadas convolucionais e as camadas completamente conectadas. A figura 3.8B mostra os *scores* de treino e de validação (fração de classificações corretas) em função do número de épocas, na qual notamos que essa rede atinge precisão ideal com somente algumas épocas de treino. A figura 3.8C mostra a matriz de confusão obtida ao aplicar a rede treinada nos 15% dos dados que não haviam sido expostos ao algoritmo anteriormente. Esses resultados indicam que nossa rede atinge precisão perfeita ao identificar a fase do cristal líquido (nemática ou isotrópica) a partir das texturas do conjunto de teste.

Também testamos se variações da arquitetura da rede mostrada na figura 3.8A são capazes de classificar a fase do cristal líquido com desempenho similar. Consideramos variações da rede quando o número de blocos de convolução n_b (seguido por camadas de agrupamento *max pooling*) muda de 1 até 5 (a rede da figura 3.8A corresponde a $n_b = 2$). Então, treinamos dez realizações de cada uma dessas redes usando o mesmo procedimento descrito para a arquitetura da figura 3.8A. Após o treinamento, estimamos a precisão média da tarefa de classificação no conjunto de teste em função de n_b. Os resultados da figura 3.8D indicam que redes com $n_b = 1, 2$ ou 3 blocos de convolução são igualmente boas em classificar as fases do cristal líquido com precisão muito próxima do valor ideal. Sendo assim, seria preferível usar $n_b = 3$ ao implantar esse modelo em uma aplicação prática, tendo em vista que o número de parâmetros ajustáveis diminui com o aumento do número de blocos de convolução, o que por sua vez simplifica os procedimentos de treinamento. Na figura 3.8D, também observamos que o desempenho da tarefa de classificação diminui significativamente quando aumentamos o número de blocos de convolução para além de $n_b = 3$, alcançando precisão de $\approx 0{,}70$ para

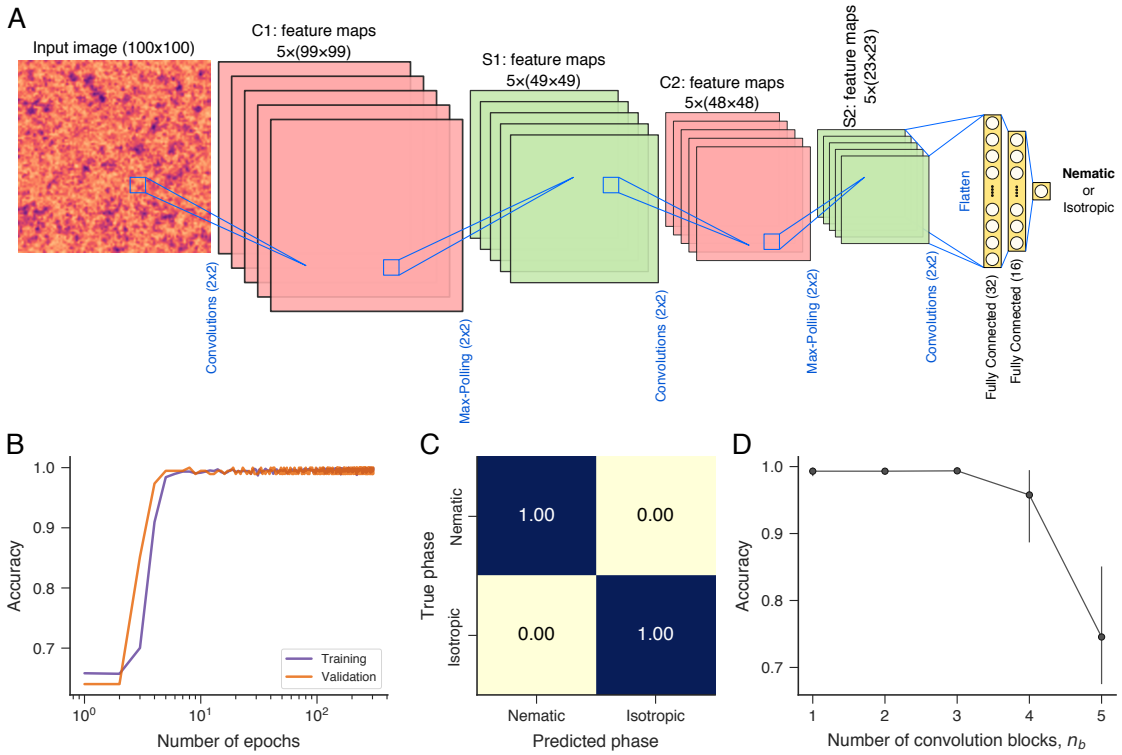

Figura 3.8: Prevendo a fase de cristais líquidos com redes convolucionais neurais. (A) Representação esquemática da arquitetura da rede utilizada para prever a fase do cristal líquido a partir de sua textura. Essa rede compreende dois blocos de camadas convolucionais (em vermelho) e de camadas de agrupamento *max pooling* (em verde), seguidas por duas camadas completamente conectadas (em amarelo) e uma camada de saída. Uma imagem de entrada de tamanho 100×100 pixels é convolvida com cinco filtros 2×2 (com passo unitário), resultando em cinco mapas de características de tamanho 99×99 (C1 em vermelho) que são passados por uma função de ativação do tipo ReLU (*rectified linear unit*). Esses mapas de características são passados para operações de agrupamento *max pooling* de tamanho 2×2 que reduzem a representação para cinco mapas de tamanho 49×49 (S1 em verde). Após esse procedimento, esses mapas são passados novamente por uma mesma configuração de blocos convolucionais e de agrupamento, resultando em cinco mapas de características de tamanho 23×23 (S2 em vermelho) que são dispostos de maneira concatenada e passam por duas camadas completamente conectadas com 32 e 16 nós, respectivamente. Finalmente, a classificação da fase (nemática ou isotrópica) ocorre na camada de saída via função de ativação sigmoide (correspondente à regressão logística). (B) *Scores* de treinamento e de validação (fração de classificações corretas) em função do número de épocas durante o estágio de treinamento. Separamos 15% dos dados como conjunto de teste e o restante (85%) é dividido em conjuntos de treinamento (80%) e validação (20%) (todos obtidos de maneira estratificada). (C) Matriz de confusão obtida ao aplicar a rede treinada no conjunto de teste. (D) Precisão da rede no conjunto de teste em função do número de blocos convolucionais e de agrupamento n_b na arquitetura (painel A corresponde a $n_b = 2$). Os círculos são os valores médios de dez realizações dos procedimentos de treino e as barras de erro representam o intervalo de confiança de 95%.

$n_b = 5$.

3.3.2 Prevendo o parâmetro de ordem de texturas nemáticas simuladas

Em uma outra aplicação, realizamos a previsão do parâmetro de ordem p de cristais líquidos simulados diretamente de suas texturas, uma tarefa análoga à apresentada na seção 3.2.1. A figura 3.9A mostra novamente a dependência de p com a temperatura T_r, na qual observamos que p diminui com o aumento de T_r. Conforme já mencionado, esse cristal líquido passa por uma transição da fase nemática para a fase isotrópica quando a temperatura excede o valor crítico $T_c = 1{,}1075$. Diferentemente da classificação da fase, agora temos um problema de regressão e, nesse caso, a saída da rede é um número contínuo representando o parâmetro de ordem p. Para essa tarefa de regressão, consideramos essencialmente a mesma arquitetura de rede usada para classificar a fase de cristais líquidos. Substituímos apenas a função de ativação sigmoide da camada de saída por uma função de ativação linear, geralmente usada em problemas de regressão. A figura 3.9B mostra a arquitetura com quatro blocos de convolução (e de agrupamento *max pooling*) $n_b = 4$.

Treinamos essa rede otimizando o erro quadrático médio (função de perda) e seguindo os mesmos procedimentos usados para a classificação da fase do cristal líquido. A figura 3.9C mostra que os coeficientes de determinação R^2 entre os valores reais e os valores preditos aproximam-se de 1 para os conjuntos de treino e validação após apenas algumas épocas de treinamento. Também, obtemos coeficiente de determinação de $\approx 0{,}997$ ao aplicar a rede treinada no conjunto de teste. Esse resultado demonstra que nossa rede convolucional neural é extremamente eficiente ao prever o parâmetro de ordem p, superando ligeiramente a abordagem descrita na seção anterior, na qual utilizamos a entropia e complexidade de permutação das imagens e o algoritmo de k-vizinhos mais próximos. A figura 3.9A mostra a comparação entre os valores previstos e reais para o parâmetro de ordem p, na qual conseguimos observar visualmente a alta precisão alcançada pela rede.

Nessa aplicação, também investigamos como o número de blocos de convolução (e de camadas de agrupamento) n_b afeta a precisão da rede. Treinamos dez realizações da rede para um dado valor de $n_b \in \{1, 2, 3, 4, 5\}$ e calculamos o valor médio do coeficiente de determinação para o conjunto de teste. A figura 3.9D mostra que essas redes apresentam excelente precisão para diferentes números de blocos de convolução, mas o desempenho ótimo ocorre para $n_b = 4$ (a arquitetura da figura 3.9B).

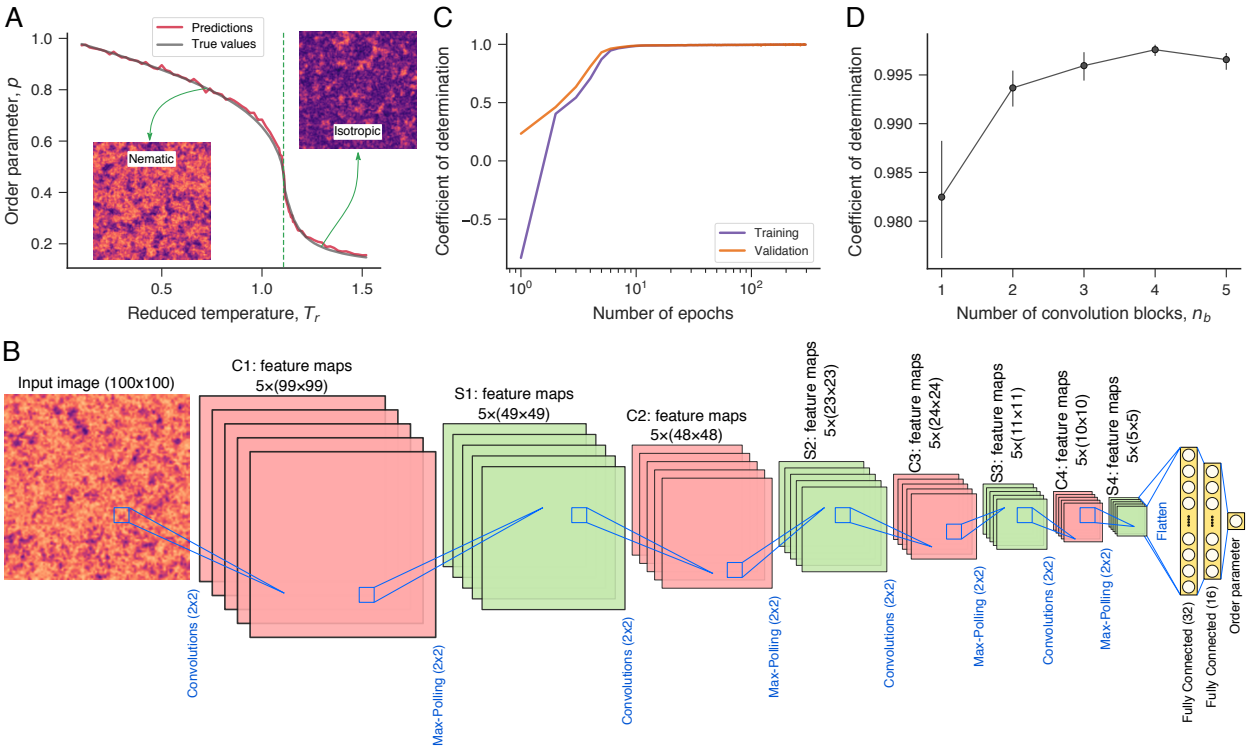

Figura 3.9: Prevendo o parâmetro de ordem de cristais líquidos com redes convolucionais neurais. (A) Dependência do parâmetro de ordem p com a temperatura reduzida T_r para um cristal liquído nemático simulado (curva cinza). A linha vertical tracejada indica a temperatura crítica $T_c = 1{,}1075$ que separa as fases nemática ($T_r < T_c$) e isotrópica ($T_r > T_c$). As inserções ilustram texturas típicas de cada fase. (B) Representação esquemática da arquitetura da rede usada para a tarefa de regressão de prever o parâmetro de ordem p a partir das texturas. Essa rede possui a mesma estrutura geral da rede usada para a classificação da fase e é composta por quatro blocos de convolução (em vermelho) e camadas de agrupamento *max pooling* (em verde) seguidos por duas camadas completamente conectadas (em amarelo) e uma camada de saída. A única diferença está na última camada, na qual usamos uma função de ativação linear para estimar o parâmetro de ordem p. (C) Coeficiente de determinação (entre os valores reais e previstos) estimado a partir dos conjuntos de treinamento e validação em função do número de épocas durante o estágio de treinamento. Separamos 15% dos dados como conjunto de teste e o restante (85%) é dividido entre os conjuntos de treino (80%) e validação (20%) obtidos de maneira estratificada. A rede treinada possui um coeficiente de determinação de $\approx 0{,}997$ quando aplicada no conjunto de teste e a curva vermelha no painel A ilustra a precisão das previsões da rede. (D) Coeficiente de determinação obtido a partir do conjunto de teste em função do número de blocos de convolução (e de agrupamento) n_b na arquitetura da rede (o painel B corresponde a $n_b = 4$). Os círculos são valores médios de dez realizações do procedimento de treinamento e as barras de erro representam o intervalo de confiança de 95%.

3.3.3 Prevendo o comprimento do passo de cristais líquidos colestéricos

Nesta seção, em uma tarefa de classificação análoga à apresentada na seção 3.2.3, vamos verificar se redes convolucionais neurais são úteis para prever o valor do passo η de cristais líquidos colestéricos diretamente das texturas desses materiais. Para isso, aplicamos basicamente a mesma arquitetura geral de rede usada com texturas nemáticas para a tarefa de classificar os valores de η. A figura 3.10A mostra a rede com $n_b = 4$ blocos de convolução (e de agrupamento *max pooling*) usada com as texturas colestéricas. Quando comparada com as redes usadas com texturas nemáticas, a única diferença está na última camada, que agora conta com 8 nós (um para cada valor do passo $\eta \in \{15, 17, 19, \ldots, 29\}$) com funções de ativação *softmax* (usualmente usadas em tarefas de classificação multiclasse). Treinamos essa rede seguindo os mesmos procedimentos usados anteriormente e considerando a entropia cruzada categórica como função de perda. A figura 3.10B mostra que os *scores* de treinamento e de validação aproximam-se da precisão ideal após aproximadamente 10 épocas de treinamento. A figura 3.10C demonstra a alta precisão dessa rede representada pela matriz de confusão calculada para o conjunto de teste (15% dos dados nunca apresentados ao algoritmo). Notamos que essa rede classifica perfeitamente todos os valores do passo. Esse desempenho é bastante superior ao apresentado na seção 3.2.3, cujo valor da precisão foi de $\approx 85\%$. Também investigamos a precisão de diferentes arquiteturas de rede alterando o número de blocos de convolução n_b. Os resultados da figura 3.10D mostram que a precisão média é bastante baixa para $n_b < 3$, atinge um valor ótimo para $n_b = 3$ e 4 e diminui quando $n_b = 5$.

3.3.4 Prevendo a temperatura de amostras experimentais

Em uma última aplicação com redes convolucionais neurais, buscamos prever a temperatura de amostras a partir de texturas experimentais de cristais líquidos E7. Essa tarefa de regressão é análoga à apresentada na seção 3.2.2. A figura 3.11A mostra exemplos de texturas obtidas em diferentes temperaturas. Inicialmente, verificamos que a arquitetura geral da rede utilizada em todas as aplicações anteriores não produz bons resultados ao lidar com essas texturas experimentais. Por causa disso, propomos uma arquitetura ligeiramente modificada ao incluir camadas adicionais de convolução antes de cada operação de agrupamento do tipo *max pooling*. Também aumentamos o número (agora são oito filtros 4×4 por bloco de convolução) bem como o tamanho dos filtros convolucionais das camadas de agrupamento (agora são 3×3 *pixels*). As camadas completamente conectadas permanecem iguais aos casos anteriores, ou seja, temos duas camadas com 32 e 16 nós, respectivamente, seguidas por uma camada de saída. Funções de ativação ReLU são usadas após todas as operações de convolução e uma função de ativação linear é usada na camada de saída. A

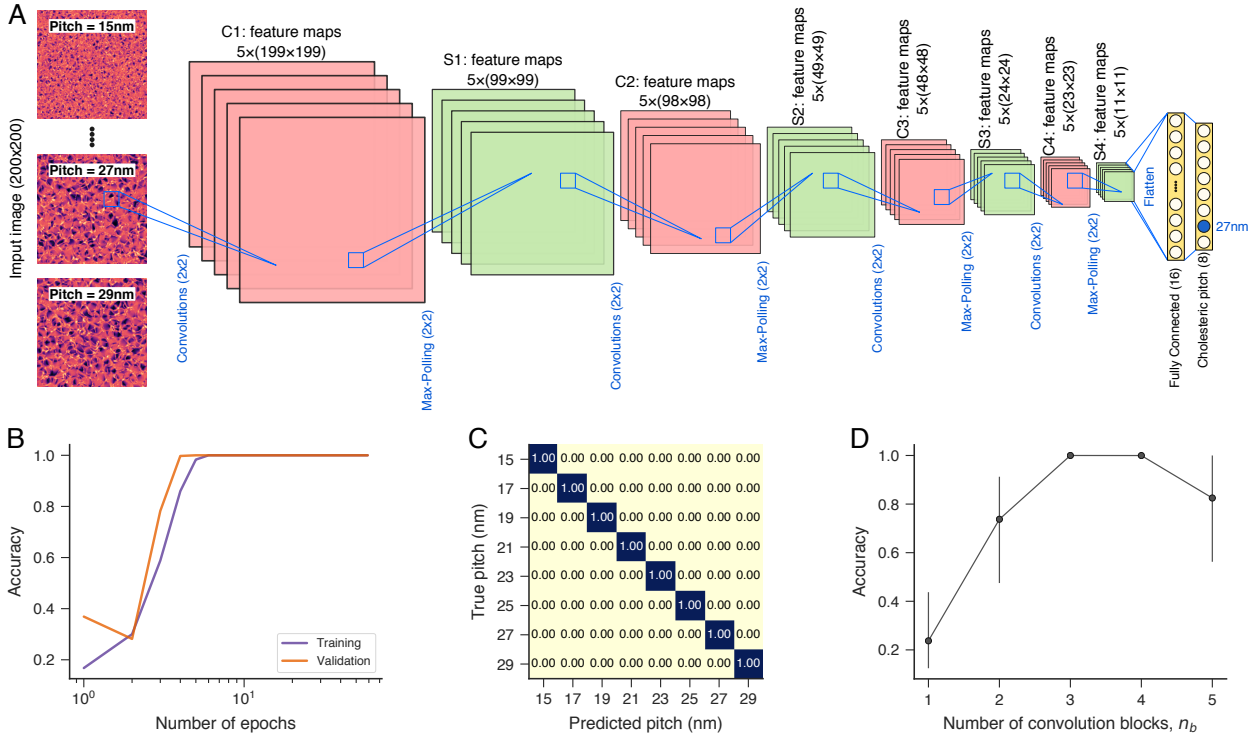

Figura 3.10: Prevendo o comprimento do passo de cristais líquidos colestéricos com redes convolucionais neurais. (A) Ilustração da arquitetura da rede usada para classificar os valores do passo. Essa rede possui a mesma estrutura geral da rede usada para classificar as fases e da rede usada para prever o parâmetro de ordem. A diferença está na última camada que agora é composta por 8 nós com funções de ativação do tipo *softmax*. (B) *Scores* de treinamento e de validação (precisão, a fração de classificações corretas) em função do número de épocas de treinamento. Separamos 15% dos dados como conjunto de teste e o restante (85%) é dividido entre os conjuntos de treino (80%) e validação (20%) obtidos de maneira estratificada. (C) Matriz de confusão obtida ao aplicar a rede treinada no conjunto de teste. A forma diagonal dessa matriz mostra que a rede treinada alcança classificação perfeita de todos os valores de passo no conjunto de teste. (D) Precisão estimada a partir do conjunto de teste em função do número de blocos de convolução (e de agrupamento) n_b na arquitetura da rede (o painel A corresponde a $n_b = 4$). Os marcadores representam os valores médios obtidos em dez realizações do procedimento de treinamento e as barras de erro representam o intervalo de confiança de 95%.

figura 3.11B ilustra a estrutura modificada da rede com $n_b = 3$ blocos de convolução.

Apesar das modificações na arquitetura da rede, os procedimentos de treinamento e de regularização permanecem os mesmos. Também utilizamos o erro quadrático médio como função de perda para esse problema de regressão. A figura 3.11C mostra o coeficiente de determinação R^2 para os conjuntos de treinamento e de validação em função do número de épocas de treinamento. Observamos que ambos *scores* aproximam-se do valor ideal após apenas algumas épocas de treinamento. Os resultados da figura 3.11D mostram a relação entre as temperaturas preditas e reais obtidas ao aplicar a rede treinada no conjunto de teste. Essa relação segue bem próxima a linha 1:1 (linha tracejada) e possui coeficiente

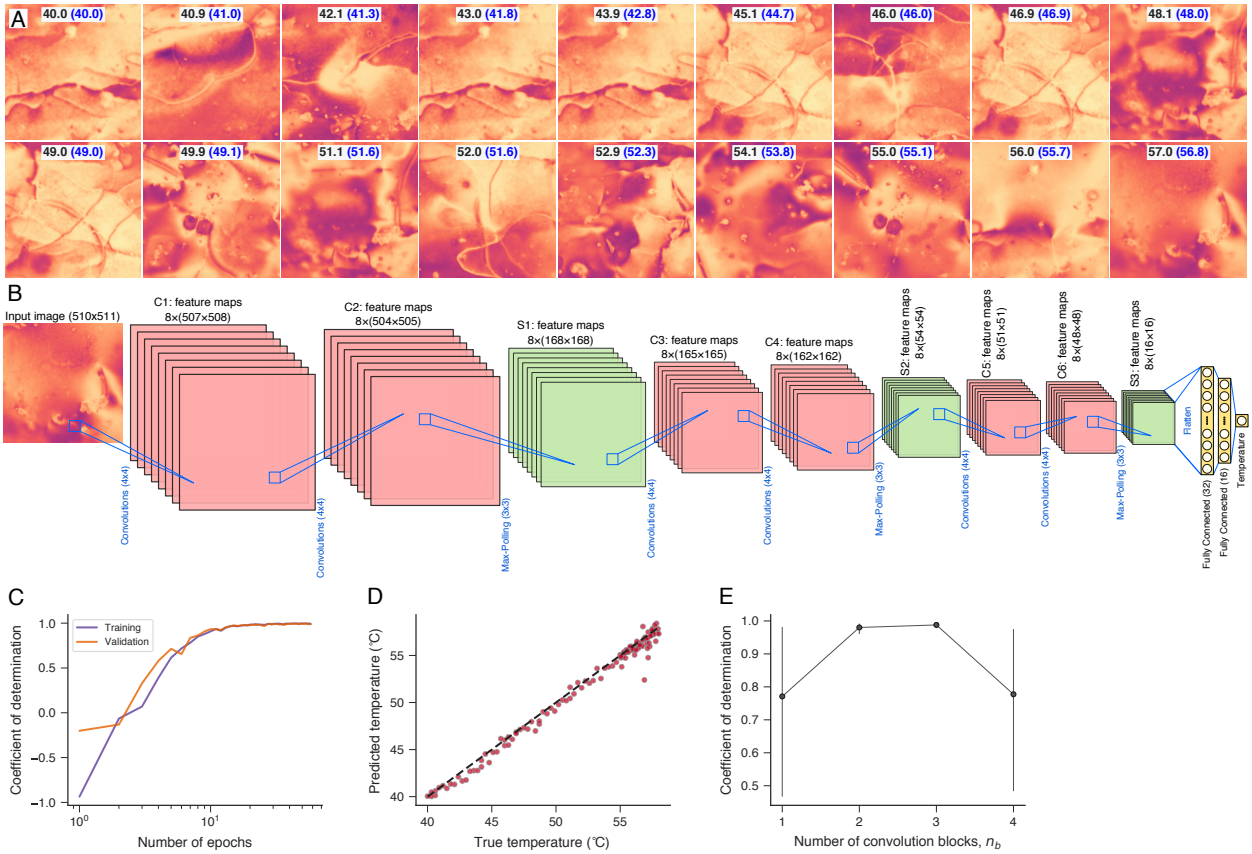

Figura 3.11: Prevendo a temperatura de amostras de cristal líquido E7 com redes convolucionais neurais. (A) Exemplos de texturas experimentais obtidas por microscopia óptica polarizada para diferentes amostras em diferentes temperaturas (indicada nas imagens em graus Celsius). (B) Representação esquemática da arquitetura da rede usada para a tarefa de regressão de prever a temperatura da amostra. Essa arquitetura de rede é ligeiramente diferente de todas as outras que usamos até agora; é composta de três blocos de oito convoluções 4×4 seguidas por outras oito convoluções 4×4 e por oito camadas de agrupamento *max pooling* 3×3. Após as últimas operações de agrupamento (S3), temos duas camadas completamente conectadas (com 32 e 16 nós, respectivamente) e uma camada de saída com um único nó e uma função de ativação linear. Todas as camadas convolucionais usam a função de ativação ReLU. (C) Coeficiente de determinação estimado a partir dos conjuntos de treinamento e de validação em função do número de épocas de treinamento. Separamos 15% dos dados como conjunto de teste e o restante (85%) é dividido entre os conjuntos de treino (80%) e validação (20%) obtidos de maneira estratificada. O coeficiente de determinação calculado para o conjunto de teste é $\approx 0{,}982$. (D) Relação entre os valores de temperatura reais e previstos obtidos ao aplicar a rede treinada no conjunto de teste (a linha tracejada representa a relação 1:1). Os valores entre parênteses no painel A também indicam as previsões da rede. (E) Coeficiente de determinação estimado a partir do conjunto de teste em função do número de blocos de convolução e de agrupamento n_b na arquitetura (o painel B corresponde a $n_b = 3$). Os marcadores representam os valores médios obtidos para dez realizações dos procedimentos de treinamento e as barras de erro representam o intervalo de confiança de 95%.

de determinação de $\approx$ 0,982, indicando a alta precisão alcançada pelo nosso método. A figura 3.11A também mostra a comparação entre os valores reais de temperatura associados com cada textura e as previsões da rede (valores entre parênteses). Vale notar também que a precisão da rede supera o desempenho da abordagem baseada na entropia e complexidade de permutação (seção 3.2.2) na qual um coeficiente de determinação de $\approx$ 0,93 foi obtido com o algoritmo de k-vizinhos mais próximos. Também investigamos a precisão da nossa abordagem com números diferentes de blocos de convolução n_b. A figura 3.11E mostra o coeficiente de determinação médio em função de n_b, na qual notamos que a precisão ótima ocorre para $n_b = 2$ ou $n_b = 3$.

3.4 Conclusão

Nesse capítulo, propusemos duas abordagens para extrair propriedades físicas de cristais líquidos diretamente de imagens das texturas desses materiais [33, 34]. Nosso primeiro método é baseado no cálculo da entropia e complexidade estatística de permutação dessas texturas, as quais são usadas como características em tarefas de aprendizagem estatística supervisionadas de regressão e classificação para prever parâmetros físicos desses materiais. Demonstramos a utilidade e precisão dessa abordagem em uma série de aplicações numéricas e experimentais. Nossos resultados mostraram que o parâmetro de ordem médio pode ser estimado a partir das imagens de texturas nemáticas obtidas por simulações de Monte Carlo com precisão de 99%. Além disso, obtivemos precisão de 92% para a tarefa de regressão de estimar a temperatura a partir de texturas experimentais de cristal líquido E7 em diferentes temperaturas e fases. Também mostramos que essa abordagem classifica diferentes comprimentos do passo de texturas colestéricas com precisão de 85%.

Nossa segunda abordagem, demonstrou a utilidade de redes convolucionais profundas em prever as mesmas propriedades físicas de cristais líquidos diretamente de suas texturas ópticas. Esse método de *deep learning* mostrou ser bastante eficiente para prever as fases (nemática ou isotrópica), parâmetros de ordem, comprimentos do passo e temperaturas das amostras de diferentes cristais líquidos. Em particular, a performance de nossas redes convolucionais neurais superou significativamente a abordagem baseada na entropia e complexidade de permutação nas tarefas associadas a determinação do passo do cristal líquido colestérico (acurácia de 100% versus 85%) e da temperatura ($R^2 = 0{,}98$ versus $R^2 = 0{,}93$).

Por um lado e além da maior precisão, os métodos baseados em redes convolucionais profundas têm a vantagem de não requererem extração manual de características das imagens. Por outro lado, vale destacar que métodos baseados em *deep learning* carecem de uma interpretação mais direta e pode ser bastante complicado associar propriedades de redes convolucionais neurais com características dos materiais em estudo. Nesse ponto, a abordagem baseada na entropia e complexidade de permutação apresenta grande vantagem. De fato,

essas medidas de complexidade são muito mais simples e intuitivas, além de serem muito rápidas e escaláveis do ponto de vista computacional.

Apesar dessas diferenças importantes, tanto os resultados baseados na entropia e complexidade de permutação, quanto aqueles obtidos via redes convolucionais neurais ajudam a reduzir a escassez de investigações com aprendizado de máquina na pesquisa de cristais líquidos. Embora métodos de aprendizado de máquina não substituam completamente os procedimentos experimentais, essas abordagens podem contribuir muito para superar várias dificuldades nas análises experimentais. A partir dos resultados mostrados nesse capítulo, esperamos que o uso de métodos de aprendizagem de máquina em situações experimentais possam se desenvolver ainda mais. Na verdade, existem vários cenários em ciência básica e aplicada que podem se beneficiar dessas técnicas, que vão muito além do uso exclusivo na pesquisa de cristais líquidos. É provável que, em um futuro próximo, técnicas de aprendizado de máquina encontrem uso em basicamente todas as ferramentas baseadas em imagens: do microscópio óptico ao eletrônico.

CAPÍTULO 4

Identificando e agrupando padrões na eficiência dos mercados de criptomoedas e de ações

Diferentemente do caráter mais aplicado do capítulo anterior, o presente capítulo retoma a abordagem do capítulo 2 de busca por padrões em sistemas complexos, com enfoque na hipótese do mercado eficiente – a ideia de que preços de ativos financeiros refletem completamente toda informação disponível. Especificamente, apresentamos dois estudos em grande escala sobre a eficiência informacional dos mercados de criptomoedas [35] e de ações [36] utilizando os valores de entropia e complexidade de permutação calculados para séries temporais financeiras oriundas desses sistemas. Em nossa primeira investigação, identificamos e agrupamos padrões na eficiência do mercado de criptomoedas. No segundo estudo, analisamos a dinâmica coletiva da eficiência dos maiores mercados de ações mundiais.

4.1 Introdução

A hipótese do mercado eficiente é um paradigma em economia e uma crença generalizada entre os agentes dos mercados de ações [152, 153]. Essa hipótese afirma que os preços dos ativos refletem completamente toda informação disponível em um mercado idealmente eficiente [154]. À medida que novas informações sobre ativos financeiros ou sobre os mercados de ações tornam-se disponíveis, essas são precificadas imediatamente pelos agentes de mercado, resultando em novos valores para os ativos [155]. Consequentemente, tentativas de prever os preços futuros em um mercado informacionalmente eficiente provavelmente não serão melhores do que uma estimativa aleatória. Portanto, a hipótese do mercado eficiente impõe limitações significativas às negociações financeiras e torna impossível tirar proveito

de mercados informacionalmente eficientes utilizando, por exemplo, operações de arbitragem ou estratégias de *trading* [155]. Além disso, acredita-se que a eficiência do mercado previne efetivamente bolhas econômicas, que estão entre as principais causas do colapso de ações e da instabilidade do mercado financeiro [156, 157]. Por outro lado, críticos da hipótese do mercado eficiente alegam que é, justamente, a crença nos mercados racionais a verdadeira culpada pela crise financeira de 2007-2008, bem como por muitos outros desenvolvimentos indesejáveis na economia mundial [158].

Se a ineficiência dos mercados financeiros é uma oportunidade de lucro ou se representa um risco sistêmico que deve ser corrigido, o fato é que a hipótese do mercado eficiente ainda é um conceito onipresente entre os agentes econômicos e acadêmicos que trabalham com dados e modelos financeiros. Esse interesse se traduz em uma grande quantidade de trabalhos que tentam quantificar o grau de eficiência de diferentes mercados de ações [159–167] e também de novas formas de investimento como as criptomoedas [168–176]. Apesar do crescente empenho em compreender melhor as diferentes facetas da hipótese do mercado eficiente, grande parte desses estudos assume que a eficiência do mercado em questão permanece inalterada ao longo do período investigado. Portanto, a possibilidade mais realista de ter mercados financeiros com grau de eficiência variável ao longo do tempo permanece muito menos explorada. Trata-se de uma limitação importante, uma vez que medidas de eficiência dependentes do tempo permitem investigar os movimentos coletivos na evolução da eficiência desses mercados, bem como quantificar a estabilidade dos *rankings* de eficiência dos mercados financeiros. Nesse sentido, apresentamos duas investigações que, dentre outras análises, definem uma eficiência variável ao longo do tempo para mercados de criptomoedas e de ações a partir de séries temporais dos retornos diários desses ativos financeiros. Como resultado comum a ambas investigações, podemos destacar que uma análise dinâmica da eficiência informacional revela padrões nos quais ativos diferentes dentro de seus mercados, ou seja, diferentes criptomoedas ou diferentes mercados de ações, são agrupados devido às similaridades entre seus perfis temporais de eficiência.

4.2 Agrupando padrões na eficiência do mercado de criptomoedas

A crescente popularidade das criptomoedas, apesar da volatilidade de seus preços, indica que o controle descentralizado por meio da tecnologia *blockchain*, juntamente com transações financeiras seguras devido à forte criptografia, são atributos altamente valorizados entre os agente financeiros de todo o mundo. Assim, uma melhor compreensão geral do mercado de moedas digitais se faz necessária. Nesse sentido, conduzimos um estudo em grande escala da eficiência desse mercado por meio da análise de séries temporais dos retornos diários dos

preços de 437 criptomoedas. Metodologicamente, nos baseamos na entropia e complexidade estatística de permutação (apresentadas no capítulo 1) calculadas em janelas de tempo móveis ao longo dessas séries temporais. Uma vez determinada a entropia e complexidade de permutação, consideramos que uma criptomoeda é informalmente eficiente em uma janela de tempo quando ambas as medidas permanecem dentro de um intervalo de confiança obtido ao embaralhar a série temporal em cada janela e calcular a entropia e complexidade de permutação para várias realizações independentes.

Como mostramos a seguir, nossa pesquisa revela que 37% das 437 criptomoedas permanecem informacionalmente eficientes em mais de 80% do tempo, enquanto 20% permanecem eficientes em menos de 20% do tempo. Notamos também que a eficiência não está correlacionada com o valor de capitalização de mercado das criptomoedas. Além disso, uma análise dinâmica da eficiência ao longo do tempo revela padrões nos quais criptomoedas diferentes são agrupadas juntas devido às similaridades de seus perfis temporais de eficiência. Para a análise dos agrupamentos, nos baseamos em uma medida de similaridade conhecida por *dynamic time warping* (DTW) [177], que possui a vantagem de determinar uma "distância" entre séries temporais de diferentes comprimentos e escalas de valores. Com base nessa análise, encontramos quatro grupos de criptomoedas: *i)* aquelas que apresentam inicialmente um nível de eficiência maior mas evoluem para um nível menor (12% do mercado); *ii)* aquelas que melhoram a eficiência ao longo do tempo (19% do mercado); *iii)* aquelas que possuem um nível de eficiência que, em média, permanece constante (43% do mercado); e aquelas que começam em um nível maior de eficiência, diminuem para um nível menor e então aumentam novamente seu nível de eficiência (26% do mercado). Essa análise dos agrupamentos também indica que as moedas mais jovens em cada grupo parecem seguir a tendência das mais velhas. Em geral, descobrimos que grande parte do mercado de criptomoedas satisfaz a hipótese do mercado eficiente na maior parte do tempo e também que o amadurecimento do mercado de moedas digitais parece ser inevitável.

4.2.1 Apresentação dos dados

Para realizar essa investigação, realizamos o *download* automatizado dos preços de fechamento diários e os valores de capitalização de 1.509 criptomoedas a partir do site `http://coinmarketcap.com` em 13 de fevereiro de 2018. Esse conjunto de dados inclui as moedas digitais mais importantes e populares que estão em circulação atualmente e abrange diferentes períodos que variam de alguns dias (por exemplo, para as criptomoedas Medicalchain e Farstcoin) até quase cinco anos (para a Bitcoin). Selecionamos as 437 criptomoedas que possuem mais de 600 observações para a análise da eficiência informacional geral. Já para a análise relacionada ao comportamento dinâmico da eficiência informacional, selecionamos as 167 criptomoedas com mais de 960 dias de observação. Esses filtros são necessários para que a estimativa da eficiência informacional em cada análise seja confiável, assegurando

que o cálculo da eficiência geral seja baseado em ao menos 100 observações da entropia e complexidade e que as séries relacionadas à eficiência informacional sejam maiores que 100 dias, conforme explicamos em detalhe a seguir.

4.2.2 Análise dos dados

Iniciamos apresentando nossa metodologia aplicada para a série temporal da Bitcoin, a criptomoeda mais conhecida e popular. Na figura 4.1A, mostramos a série temporal do retorno logarítmico R_t dos preços de fechamento diários para essa moeda, definida por

$$R_t = \log P_t - \log P_{t-1}, \tag{4.1}$$

com $\log P_t$ e $\log P_{t-1}$ representando os logaritmos naturais dos preços de fechamento nos tempos t e $t-1$. Essa série temporal (de comprimento n) é amostrada utilizando uma janela móvel (sombreada em cinza na figura 4.1A) que contém 500 pontos, o que corresponde a aproximadamente dois anos de atividade econômica. Essa janela móvel tem passo diário, ou seja, move-se adiante um dia de cada vez e, para cada passo, determinamos a entropia de permutação H_t e a complexidade estatística C_t (como definidas no capítulo 1) usando a *embedding dimension* $d = 4$.

Esse procedimento define novas séries temporais que representam os valores de H_t e C_t em cada janela, com t sendo a data central das janelas, como é mostrado nas figuras 4.1B e 4.1C. Utilizamos janelas de 500 dias para satisfazer a condição $d! \ll n$ e, assim, obtermos uma estimativa confiável para a entropia H e a complexidade estatística C de permutação. Conforme discutimos no capítulo 1, bem como nas investigações reportadas nos dois capítulos anteriores, essas duas medidas de complexidade são estimadas a partir dos padrões de ordem local entre valores consecutivos de R_t. Sendo assim, a entropia de permutação mede o grau de aleatoriedade na ocorrência de padrões ordinais e varia de $H \approx 0$ (para uma série completamente regular) a $H \approx 1$ (para uma série completamente aleatória). A complexidade estatística C, por sua vez, quantifica a complexidade estrutural nessa dinâmica de ordenamento. Nesse caso, temos $C \approx 0$ em ambos os extremos de ordem e desordem, enquanto $C > 0$ indica que os padrões ordinais ocorrem de uma maneira mais complexa.

Consideramos que uma criptomoeda adere à hipótese de mercado eficiente quando os valores de H_t e C_t dentro de uma janela temporal não podem ser distinguidos daqueles obtidos ao acaso. Para determinar a condição anterior, calculamos um intervalo de confiança de 95% (região sombreada em cinza nas figuras 4.1B e C) embaralhando os dados em cada janela temporal e calculando H e C para 100 realizações independentes. Dessa maneira, definimos a eficiência informacional geral E de uma determinada criptomoeda como sendo a fração de tempo em que os valores de entropia e complexidade permanecem simultaneamente dentro do intervalo de confiança.

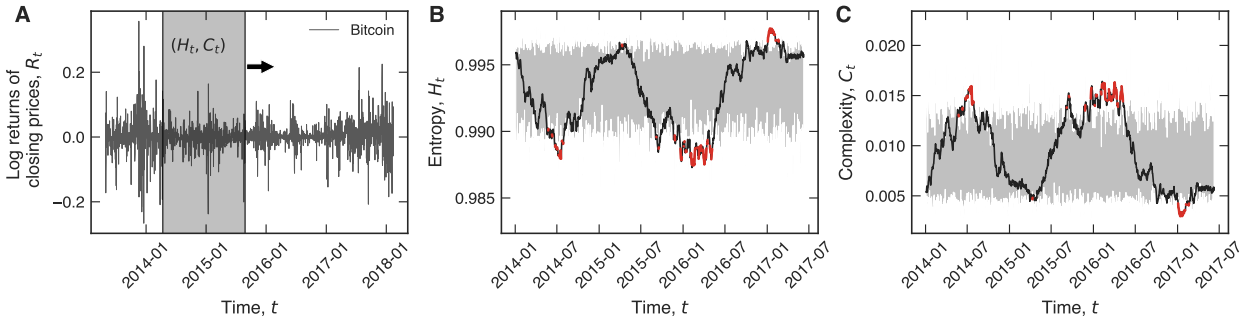

Figura 4.1: Quantificando a eficiência informacional de cripto-ativos com a entropia de permutação e a complexidade estatística. (A) Série temporal dos retornos logarítmicos R_t dos preços de fechamento da Bitcoin de 28 de abril de 2013 a 13 de fevereiro de 2018. A área sombreada representa uma janela móvel de 500 dias (aproximadamente dois anos de atividade econômica) que move-se adiante um dia de cada vez. As curvas em preto nos painéis (B) e (C) mostram a evolução temporal da entropia de permutação H_t e da complexidade estatística C_t calculadas dentro da janela móvel, respectivamente. A *embedding dimension* utilizada foi $d = 4$. As áreas sombreadas representam intervalos de confiança de 95% obtidos ao embaralhar os dados em cada janela e calcular os valores da entropia e da complexidade para várias realizações independentes. Os segmentos vermelhos indicam os valores de H_t e C_t que estão fora do intervalo de confiança. Definimos a eficiência informacional geral E como a fração de tempo (dias) em que ambas as medidas de complexidade permanecem dentro do intervalo de confiança de 95%. A eficiência estimada para a Bitcoin é $E \approx 0{,}85$ no período em análise.

Portanto, os valores de E variam entre zero e um, sendo que o limite inferior indica uma criptomoeda com eficiência muito baixa, enquanto o limite superior representa uma criptomoeda altamente eficiente. É interessante notar que essa definição está em bom acordo com as ideias fundamentais subjacentes à hipótese do mercado eficiente, no sentido em que o preço de um cripto-ativo com um valor alto para E deve ser bastante robusto contra estratégias lucrativas de negociação. Por outro lado, o preço de uma criptomoeda que possui um valor baixo para E é mais provável de ser previsto e vulnerável à essas estratégias. Para o exemplo da figura 4.1, isto é, o mercado da criptomoeda Bitcoin, encontramos $E \approx 0{,}85$, indicando que esse cripto-ativo é aderente à hipótese do mercado eficiente, em linha com o que foi reportado em pesquisas anteriores [172,173]. Esse fato valida nossa abordagem e nos convida a uma análise em grande escala nos mesmos moldes.

Para isso, calculamos os valores de E para as 437 criptomoedas do nosso conjunto de dados e estimamos a sua distribuição de probabilidade, conforme mostra a figura 4.2A. Podemos observar que essa distribuição possui uma forma bimodal bem pronunciada, de tal modo que o primeiro pico contém $\approx 20\%$ das criptomoedas, enquanto o segundo contém $\approx 37\%$. O primeiro pico corresponde às moedas digitais mais informacionalmente ineficientes, já o segundo representa as mais eficientes. Evidentemente, existem significativamente mais moedas informacionalmente eficientes do que ineficientes no mercado. No entanto, a afirmação

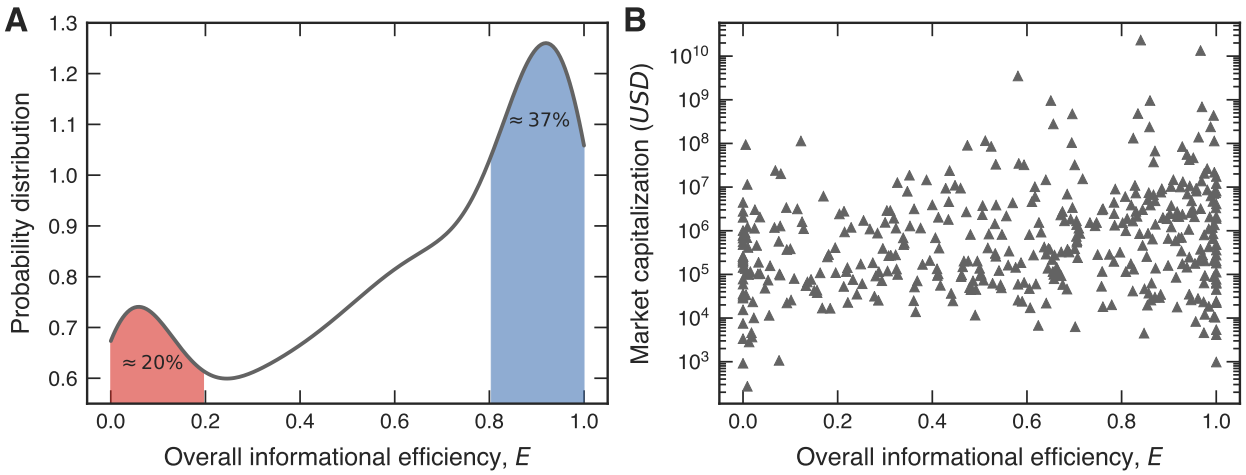

Figura 4.2: Eficiência geral do mercado de criptomoedas e o desacoplamento do valor médio de capitalização de mercado. (A) Estimativa de densidade de *kernel* da função distribuição de probabilidade da eficiência informacional E para todas as 437 criptomoedas que possuem mais de 600 observações de R_t. Notamos que essa distribuição é bimodal. O primeiro pico (para E menor que 0,2) contém $\approx 20\%$ das criptomoedas (área sombreada vermelha), indicando os casos mais informacionalmente ineficientes. Já no segundo pico (para E maior que 0,8) temos $\approx 37\%$ das criptomoedas (área sombreada azul), as mais eficientes. (B) Gráfico de dispersão para os valores da eficiência informacional E versus o valor médio de capitalização de mercado em representação linear-log. Não observamos correlação entre essas variáveis, indicando um desacoplamento entre a eficiência informacional e o valor médio de capitalização do mercado. O coeficiente de correlação linear de Pearson é $\approx 0{,}07$ com p-valor $\approx 0{,}16$, indicando que a hipótese nula de não existir correlação linear não pode ser rejeitada.

de que temos um mercado completamente aderente à hipótese do mercado eficiente ainda é exagerada.

Na figura 4.2B, mostramos como a média do valor de capitalização do mercado depende da eficiência informacional E. Notamos que não há quase nenhuma correlação entre o valor médio de capitalização e a eficiência informacional. De fato, uma análise estatística revela que o coeficiente de correlação linear de Pearson é $\approx 0{,}07$, com p-valor $\approx 0{,}16$. Consequentemente, é impossível rejeitar a hipótese nula de que não existe correlação linear entre essas duas quantidades. Como o valor de capitalização determina o valor de mercado da criptomoeda (isto é, o número de criptomoedas multiplicado pelo preço atual de mercado), encontramos que a aderência do mercado de criptomoedas à hipótese de mercado eficiente não depende da volatilidade dos preços em anos recentes.

Agora vamos nos concentrar em uma visão mais detalhada do comportamento dinâmico da eficiência informacional. Para isso, selecionamos as 167 criptomoedas que possuem mais de 460 observações para H_t e C_t. A partir dessas séries temporais, definimos a eficiência informacional dependente do tempo E_t utilizando uma janela móvel de 360 pontos sobre as séries temporais de H_t e C_t. Essa janela desloca-se adiante um dia de cada vez e o valor

de E_t é a fração de dias na qual H_t e C_t contidos na janela permanecem simultaneamente dentro do intervalo de confiança de 95%.

Essas novas séries temporais E_t fornecem uma visão detalhada sobre como a eficiência informacional das criptomoedas muda ao longo do tempo, o que, por sua vez, nos permite encontrar as moedas que possuem características temporais similares. A figura 4.3A mostra três exemplos da evolução temporal da eficiência informacional E_t para as moedas BitcoinDark, 42-coin e Diamond. Essa pequena amostra já indica que a forma de E_t pode ser bem similar entre algumas criptomoedas, como para a BitcoinDark e a Diamond.

Para estender essa análise comparativa para todas as 167 criptomoedas, utilizamos o algoritmo *dynamic time warping* [178] (DTW) para medir a similaridade entre todos os pares possíveis de séries temporais E_t. A DTW é uma medida de similaridade baseada na forma da série, a qual tem como principal vantagem a possibilidade de ser aplicada em séries temporais de tamanhos e escalas de valores diferentes. Esse método calcula um alinhamento ótimo entre duas séries temporais minimizando uma função custo (ou a distância) [177] e é amplamente empregado para tarefas de agrupamento de séries temporais [179]. Para os exemplos da figura 4.3A, a distância DTW entre BitcoinDark e Diamond é 3,64, enquanto as distâncias entre 42-coin e essas duas outras moedas são 7,73 e 11,72, respectivamente. Assim, quanto menor for a distância DTW entre um par de criptomoedas, mais similar será o perfil da evolução de E_t entre elas. A figura 4.3B mostra o gráfico da matriz das distâncias DTW entre todos os pares das 167 criptomoedas selecionadas.

Similarmente ao que fizemos para os estilos artísticos de obras de arte no capítulo 2, investigamos uma possível organização hierárquica das criptomoedas com relação a evolução da eficiência informacional. Para isso, utilizamos um critério de agrupamento que é baseado na distância média entre pares de séries das moedas, conhecido por *average linkage criteria*, para construir uma representação em dendrograma da matriz das distâncias DTW. Esse procedimento agrupa, iterativamente, pares de grupos que possuem a menor distância média, a qual é definida como o valor médio da distância entre todos os pares de elementos dos dois grupos. A figura 4.3C mostra que esse dendrograma corrobora a ideia de que as criptomoedas são hierarquicamente organizadas com relação ao perfil da evolução da eficiência informacional E_t.

Essa organização hierárquica também nos permite encontrar os grupos de moedas com características temporais similares. Para isso, precisamos encontrar uma distância limiar ótima para segmentar o dendrograma e dividir as criptomoedas em grupos. Conforme vimos no capítulo 2, uma abordagem natural para especificar essa distância limiar é maximizar o coeficiente de silhueta, que quantifica a consistência do procedimento de agrupamento. Maximizando o coeficiente de silhueta, encontramos que 5,7 é a distância DTW ótima que divide o dendrograma nos quatro grupos de criptomoedas indicados pelos ramos coloridos na figura 4.3C. Dessa maneira, criptomoedas que pertencem ao mesmo grupo possuem per-

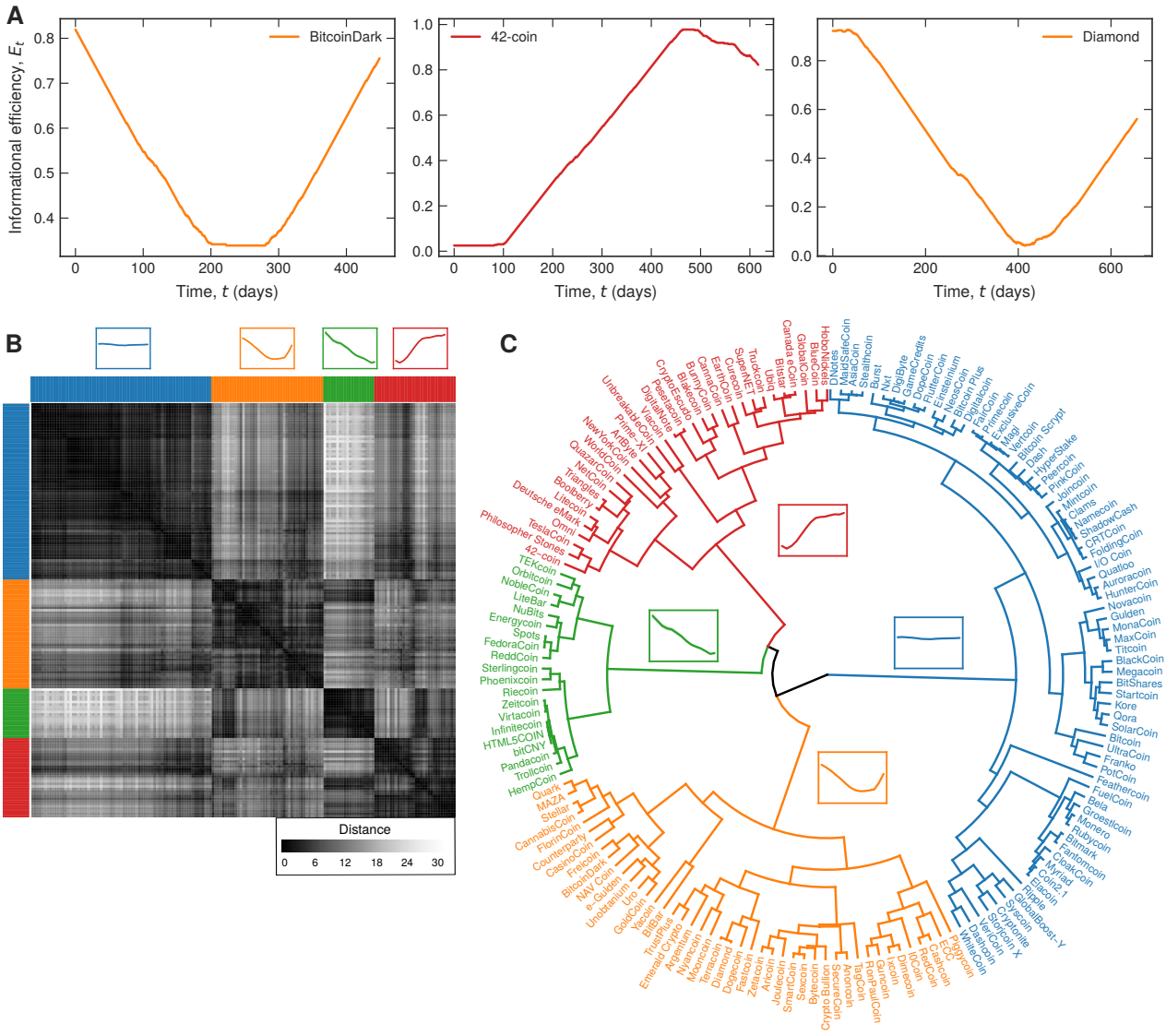

Figura 4.3: Agrupamento de padrões na dinâmica da eficiência informacional. (A) Três exemplos da evolução temporal da eficiência informacional E_t. A dinâmica de E_t é obtida usando uma janela móvel de 360 dias sobre as séries temporais H_t e C_t e calculando a fração de pontos (dias) que ambas as medidas de complexidade permanecem dentro do intervalo de confiança de 95%. Notamos que os perfis de E_t são muito similares para BitcoinDark e Diamond, enquanto 42-coin mostra uma trajetória diferente para E_t. (B) Gráfico da matriz das distâncias obtidas pelo algoritmo *dynamic time warping* (DTW) entre todos os pares das 167 criptomoedas que possuem mais de 460 observações para H_t e C_t. (C) Dendrograma mostrando o resultado do procedimento de agrupamento hierárquico baseado nas distâncias DTW utilizando o critério *average linkage* [71]. Os ramos coloridos indicam os quatro grupos de criptomoedas que exibem padrões similares para a dinâmica de E_t. Esses grupos são obtidos segmentando o dendrograma na distância limiar que maximiza o coeficiente de silhueta. A ordem das linhas e das colunas no gráfico da matriz (B) é a mesma usada no dendrograma e as inserções coloridas desse painel representam a tendência média de E_t para cada grupo.

fis temporais bem similares para a eficiência informacional E_t, cujo comportamento médio corresponde aos observados nas quatro inserções no topo da figura 4.3B. Esses quatro perfis médios podem ser qualitativamente descritos por: *i*) um nível de eficiência alto e, em média, constante (grupo azul, 43% das criptomoedas); *ii*) uma eficiência informacional que começa em um nível maior, decresce para um menor e aumenta novamente (grupo laranja, 26% das criptomoedas); *iii*) uma eficiência informacional decrescente (grupo verde, 12% das criptomoedas); e *iv*) uma eficiência informacional crescente (grupo vermelho, 19% das criptomoedas).

Por fim, é informativo analisar a evolução temporal da eficiência informacional E_t para cada um dos quatro grupos, de tal maneira que moedas com idades diferentes são distinguidas, como mostrado na figura 4.4. Observamos, nos quatro grupos, que as tendências das moedas pioneiras parecem ser seguidas pelas moedas criadas mais recentemente. Também é interessante notar que 81% das criptomoedas pertencem aos grupos caracterizados por uma eficiência informacional alta e constante ou por um nível crescente de eficiência. Isso sugere que as moedas mais jovens, que atualmente não aderem à hipótese do mercado eficiente, podem muito bem fazê-lo em um futuro próximo. Desse modo, o mercado de moedas digitais pode se tornar tão compatível com a hipótese de mercado eficiente quanto os mercados financeiros tradicionais geralmente são.

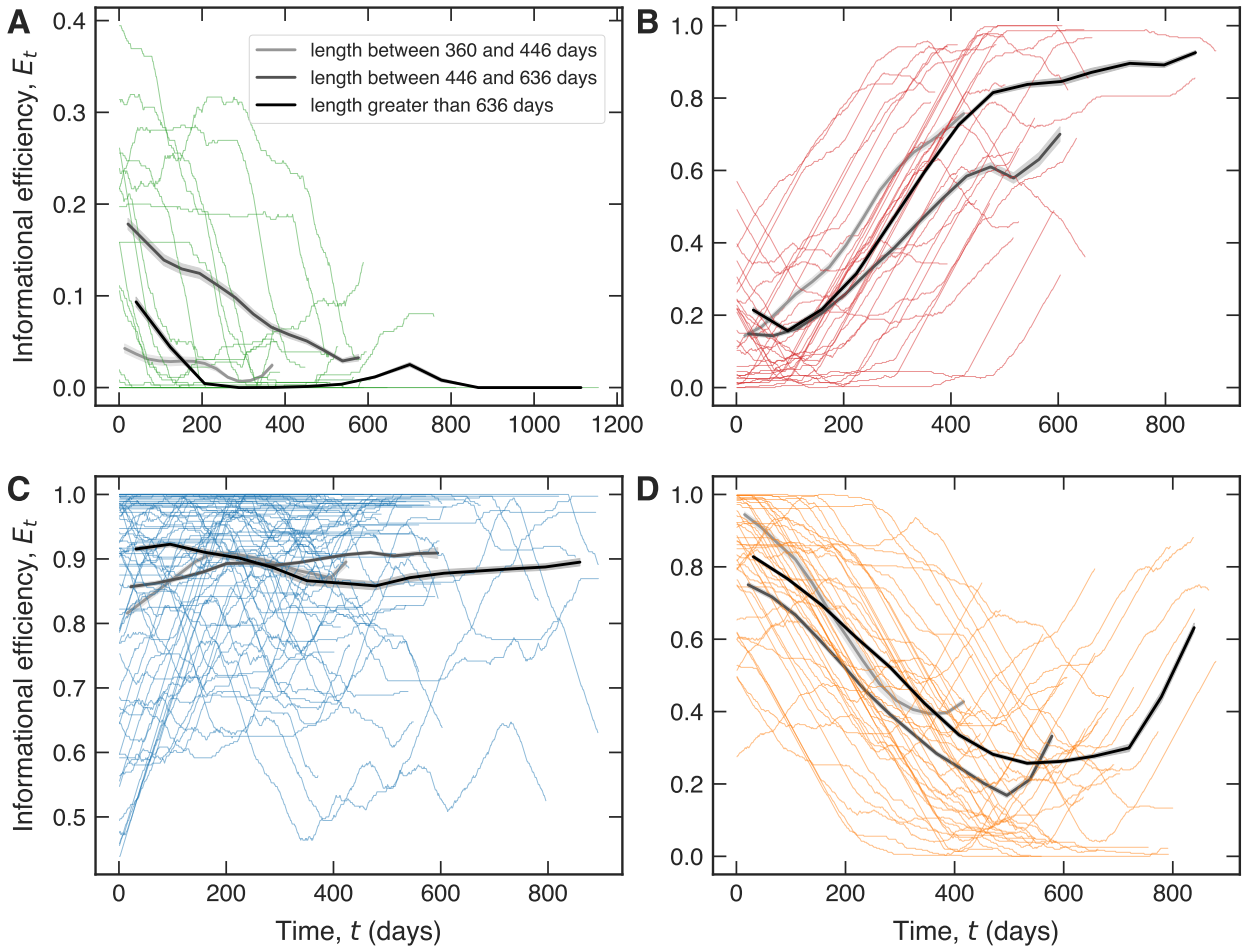

Figura 4.4: Evolução temporal da eficiência informacional dos quatro grupos diferentes de criptomoedas. Os painéis (A) ao (D) mostram séries temporais da eficiência informacional E_t para os quatro grupos de criptomoedas obtidos pelo procedimento de agrupamento hierárquico. As curvas coloridas representam cada uma das 167 criptomoedas, enquanto as curvas cinza mostram o comportamento médio de E_t em cada grupo para três intervalos de comprimento dessas séries. Esses intervalos de comprimento foram escolhidos para conter aproximadamente o mesmo número de criptomoedas. Em (A), observamos predominantemente um padrão decrescente para E_t. Em (B), temos a prevalência de uma tendência crescente para as criptomoedas mais jovens até um ponto em que a eficiência começa a saturar, em particular para as criptomoedas mais velhas. Em (C), o comportamento de E_t varia ao longo do tempo e, em média, é aproximadamente constante e altamente eficiente. Finalmente, em (D), observamos uma tendência decrescente inicial, seguida por um crescimento no fim do período de observação. Esses resultados sugerem que as criptomoedas mais jovens de cada grupo provavelmente seguirão a tendência das moedas pioneiras.

4.3 Dinâmica coletiva da eficiência do mercado de ações

Embora a hipótese do mercado eficiente seja um conceito crucial em economia, os mercados de ações reais não são idealmente eficientes o tempo todo [157], assim como também observamos na seção anterior para o mercado de criptomoedas. Os preços das ações em mercados reais podem tornar-se autocorrelacionados durante períodos de curto prazo [180], corroborando a ideia mais holística de que realizar previsões de curto prazo e arbitragem são possíveis mesmo nos mercados mais eficientes. Outra evidência de que os mercados de ações não são idealmente eficientes são as flutuações não Gaussianas (distribuições de cauda longa) dos retornos logarítmicos dos preços dos ativos [181, 182], a dificuldade de modelos simples de caminhada aleatória em prever a quebra dos mercados de ações [183], e a existência de estratégias de *trading* bem-sucedidas [184]. Além disso, embora acredita-se que mercados eficientes previnem bolhas econômicas e quebras do mercado [156], são justamente as correlações de longo alcance presentes em tais eventos que os tornam mais previsíveis [185].

No estudo apresentado nessa seção, investigamos o comportamento dinâmico da eficiência de 43 mercados de ações mundiais durante os últimos 20 anos. Utilizamos uma abordagem similar a usada anteriormente com criptomoedas para definir a eficiência variável no tempo a partir dos retornos logarítmicos dos índices dos mercados de ações. Especificamente, definimos o grau de eficiência dependente do tempo de um mercado de ações como a entropia de permutação [28] calculada dentro de uma janela móvel dos retornos logarítmicos. Nesse caso, não utilizamos a comparação com séries embaralhadas, pois nosso enfoque com o mercado de ações é diferente. No lugar de testar a hipótese de mercado eficiente, estamos agora interessados em analisar a evolução coletiva do grau de eficiência desses mercados, ou seja, até que ponto variações na eficiência de um mercado afetam a eficiência de outros mercados. Além disso, não usamos a complexidade estatística para essa análise, pois essa medida mostrou-se muito correlacionada com a entropia de permutação e, portanto, acabaria por fornecer a mesma informação sobre a evolução coletiva da eficiência dos mercados de ações.

Nossa pesquisa mostra que os principais mercados de ações podem ser hierarquicamente classificados em vários grupos de acordo com a similaridade na evolução de seus graus de eficiência de longo prazo. No entanto, descobrimos que os *rankings* de eficiência dos mercados de ações e grupos de mercados com tendências de eficiência similares são estáveis apenas em curtos períodos de tempo que, geralmente, não se estendem por mais do que alguns meses. Esse resultado indica que agrupar os mercados de acordo com o comportamento de longo prazo pode ocultar aspectos importantes de suas interações. Para revelar essas características, utilizamos uma abordagem dinâmica de agrupamento em que é possível identificar grupos de mercados de ações com padrões de eficiência semelhantes dentro de uma janela de tempo. Utilizando esses grupos que variam ao longo do tempo, construímos uma rede com-

plexa ponderada na qual nós representam mercados de ações, conexões indicam mercados que aparecem juntos no mesmo grupo pelo menos uma vez e os pesos das conexões são proporcionais ao número de vezes que um par de mercados aparece no mesmo grupo. Essa rede complexa permite identificar os mercados mais influentes, bem como sua estrutura modular, que consiste em dois grupos distintos de mercados com tendências de eficiência similares. Observamos ainda que a rede de mercados de ações é bastante densa e emaranhada. Assim, a dinâmica da eficiência do mercado de ações parece ser um fenômeno coletivo que pode fazer com que todo o sistema financeiro opere em um nível muito alto de eficiência informacional, mas que também coloca todo o sistema sob risco contínuo e sistêmico de falha.

4.3.1 Apresentação dos dados

O conjunto de dados usado nesse estudo foi obtido a partir da *API* de dados históricos financeiros do *Yahoo!* (via módulo Python *yfinance* [186]), e também a partir dos dados de mercado do *Wall Street Journal* [187] e do site *investing.com* (via módulo Python *investpy* [188]). Primeiro, obtivemos os códigos dos 43 índices dos maiores mercados de ações (tabela 4.1) e, em seguida, fizemos o *download* dos preços de fechamento diário ajustados de cada mercado a partir do *Yahoo! finance* no período de 1º de janeiro de 2000 a 31 de outubro de 2020 (cada série temporal possui 5204 pontos). Os mercados que não possuem dados na base do *Yahoo! finance* ou que estavam incompletos, foram obtidos a partir do *Wall Street Journal*. Para os mercados que, mesmo assim, ainda estavam incompletos ou indisponíveis nas bases de dados anteriores, nós obtivemos as séries históricas a partir do site `http://investing.com`. A tabela 4.1 mostra uma lista dos índices de todos os mercados usados em nossa investigação.

4.3.2 Análise dos dados

A partir das séries temporais dos preços de fechamento, calculamos os retornos logarítmicos $R(t)$ (tal como fizemos na seção 4.2.2 para o preço das criptomoedas) de cada índice de mercado, no qual t representa a data de fechamento. A figura 4.5A ilustra a evolução temporal de $R(t)$ para o índice S&P 500 (um indicador influente do mercado de ações dos EUA). Em seguida, amostramos a série dos retornos logarítmicos com uma janela móvel de 500 dias (área sombreada na figura 4.5A), que corresponde aproximadamente a 2 anos de atividade econômica. A janela móvel possui passo diário e, para cada janela, calculamos a entropia de permutação normalizada $H(t)$ [28] com *embedding dimension* $d = 4$. Esse procedimento cria uma série temporal da entropia de permutação $H(t)$ para cada mercado de ações, como é mostrado na figura 4.5B para o índice S&P 500 (veja a figura 4.6 para todos os mercados).

Conforme definido no capítulo 1 e também discutido na seção 4.2.2, a entropia de permu-

	Índice do mercado	Localização	Código	Fonte de dados
1	MERVAL	Argentina	^MERV	Yahoo finance
2	ALL ORDINARIES	Austrália	^AORD	Yahoo finance
3	S&P/ASX 200	Austrália	^AXJO	Yahoo finance
4	ATX Index	Áustria	^ATX	Yahoo finance
5	BEL 20	Bélgica	^BFX	Yahoo finance
6	IBOVESPA	Brasil	^BVSP	Yahoo finance
7	S&P/TSX Composite index	Canadá	^GSPTSE	Yahoo finance
8	SSE Composite Index	China	000001.SS	Yahoo finance
9	Shenzhen Component	China	399001.SZ	Yahoo finance
10	EURONEXT 100	Europa	^N100	Yahoo finance
11	STOXX Europe 50 Index	Europa	^STOXX50E	Wall Street Journal
12	CAC 40	França	^FCHI	Yahoo finance
13	DAX PERFORMANCE-INDEX	Alemanha	^GDAXI	Yahoo finance
14	HANG SENG INDEX	Hong Kong	^HSI	Yahoo finance
15	S&P BSE SENSEX	Índia	^BSESN	Yahoo finance
16	Jakarta Composite Index	Indonésia	^JKSE	Yahoo finance
17	TA-35 Index	Israel	TA35.TA	Yahoo finance
18	FTSE Italia All-Share Index	Itália	FTSEMIB.MI	Yahoo finance
19	Nikkei 225	Japão	^N225	Yahoo finance
20	FTSE Bursa Malaysia KLCI	Malásia	^KLSE	Yahoo finance
21	IPC MEXICO	México	^MXX	Yahoo finance
22	Amsterdam AEX Index	Países Baixos	^AEX	Yahoo finance
23	Oslo Bors All Share Index	Noruega	^OSEAX	Wall Street Journal
24	Pakistan Stock Exchange	Paquistão	^KSE	Yahoo finance
25	PSEi Index	Filipinas	PSEI.PS	Yahoo finance
26	WIG20	Polônia	WIG20	investing.com
27	MOEX Russia Index	Rússia	IMOEX.ME	investing.com
28	Russian Trading System (RTS) Index	Rússia	RTSI.ME	Wall Street Journal
29	Tadawul All Shares Index	Árabia Saudita	^TASI.SR	investing.com
30	STI Index	Singapura	^STI	Yahoo finance
31	Top 40 USD Net TRI Index	África do Sul	^JN0U.JO	Wall Street Journal
32	KOSPI Composite Index	Coréia do Sul	^KS11	Yahoo finance
33	IBEX 35 Index	Espanha	^IBEX	Yahoo finance
34	Swiss Market Index	Suíça	^SSMI	Yahoo finance
35	TSEC weighted index	Taiwan	^TWII	Yahoo finance
36	SET Index	Tailândia	^SET.BK	investing.com
37	FTSE 100	Reino Unido	^FTSE	Yahoo finance
38	Dow 30	EUA	^DJI	Yahoo finance
39	NYSE AMEX COMPOSITE INDEX	EUA	^XAX	Yahoo finance
40	NYSE COMPOSITE (DJ)	EUA	^NYA	Yahoo finance
41	Nasdaq	EUA	^IXIC	Yahoo finance
42	Russell 2000	EUA	^RUT	Yahoo finance
43	S&P 500	EUA	^GSPC	Yahoo finance

Tabela 4.1: Os 43 maiores mercados de ações mundiais usados em nosso estudo.

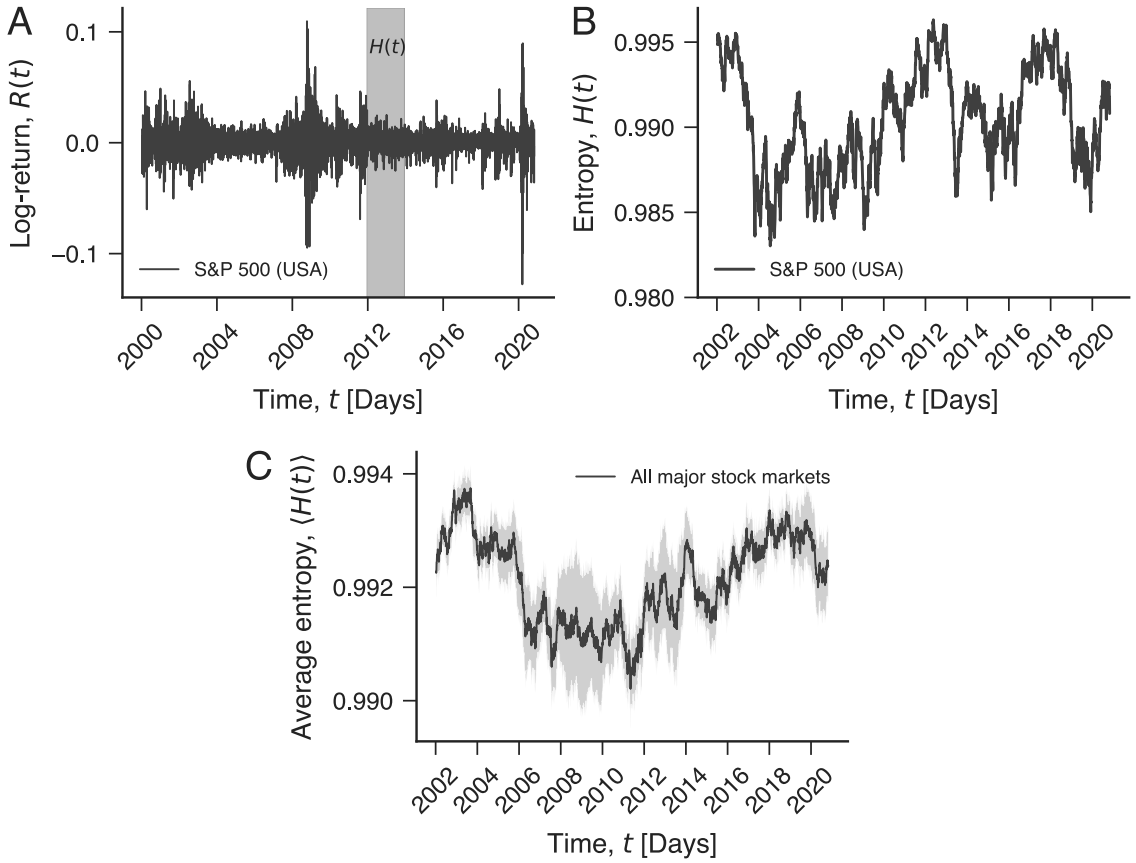

Figura 4.5: Definindo a eficiência informacional dos mercados de ações com a entropia de permutação. (A) Série temporal dos retornos logarítmicos $R(t)$ dos preços de fechamento diários do índice S&P 500 de 1º de janeiro de 2000 a 31 de outubro de 2020. A área sombreada ilustra uma janela móvel de 500 dias (dois anos de operação do mercado de ações) usada para calcular a entropia de permutação $H(t)$. (B) Evolução temporal da entropia de permutação $H(t)$ com *embedding dimension* $d = 4$ do índice S&P 500. (C) Evolução temporal do valor médio da eficiência $\langle H(t) \rangle$ de todos os 43 mercados de ações presentes em nosso estudo (a área sombreada representa o erro padrão da média).

tação é estimada a partir dos padrões ordinais entre valores consecutivos de $R(t)$ e quantifica o grau de aleatoriedade na ocorrência desses padrões. Esperamos que uma série completamente regular possua $H \approx 0$, enquanto uma série completamente aleatória tenha $H \approx 1$. Assim, quanto maior o valor de $H(t)$, mais aleatória é a série dos retornos logarítmicos por volta do tempo t e mais informacionalmente eficiente é o mercado de ações naquele momento específico. Por outro lado, uma diminuição em $H(t)$ indica a emergência de um comportamento mais regular (e possivelmente mais previsível) da série dos retornos logarítmicos e, portanto, um período menos eficiente do mercado de ações. Também estimamos o comportamento médio da eficiência $\langle H(t) \rangle$ para todos os mercados de ações. A figura 4.5C mostra que o comportamento agregado é mais suave do que o comportamento observado para mercados individuais e parece refletir grandes eventos financeiros, tais como a "crise financeira" global (2007-2008), já que por volta desse período $\langle H(t) \rangle$ possui valores menores de entropia.

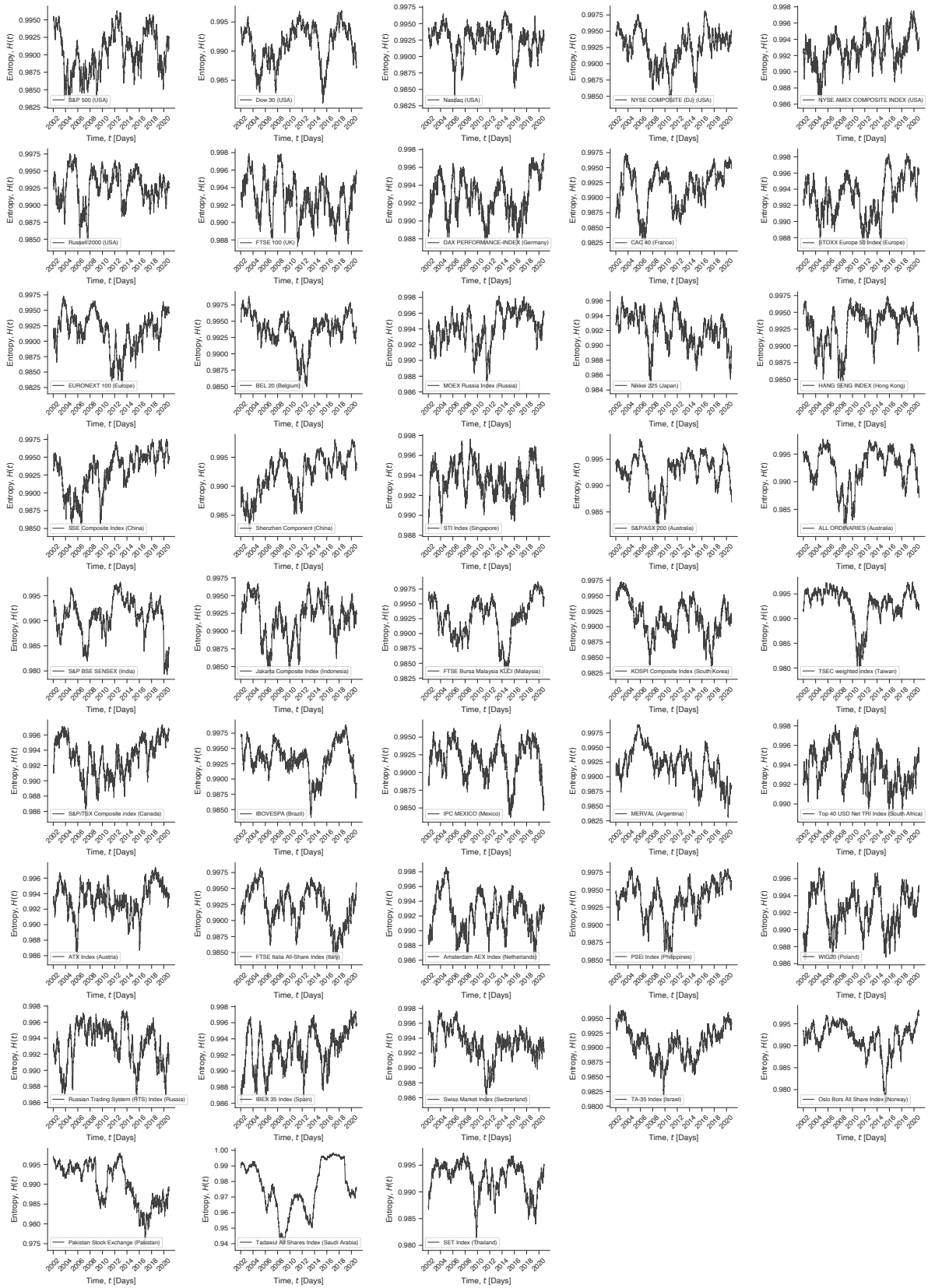

Figura 4.6: Evolução temporal da entropia de permutação $H(t)$ com *embedding dimension* $d = 4$ para os 43 mercados de ações em nosso estudo.

O comportamento de investidores tende a sincronizar durante as quedas dos mercados de ações [157], e estratégias de investimentos podem propagar choques por meio da rede financeira e levar ao surgimento de fortes correlações entre mercados financeiros [189]. Da mesma forma, esperamos que esses comportamentos coletivos afetem a dinâmica da eficiência do mercado de ações e produzam movimentos conjuntos em $H(t)$, capazes de organizar os mercados em estruturas hierárquicas com tendências de eficiência similares. Para investigar essa possibilidade, primeiro estimamos a distância de correlação das séries temporais da eficiência entre todos os pares de mercados, criando a matriz de distâncias das correlações definida por

$$d(H_i, H_j) = \sqrt{2(1 - \rho(H_i, H_j))}, \tag{4.2}$$

na qual $\rho(H_i, H_j)$ é o coeficiente de correlação de Pearson entre as séries temporais H_i do i-ésimo mercado de ações e H_j a mesma quantidade para o j-ésimo mercado. Essa matriz é mostrada na figura 4.7A.

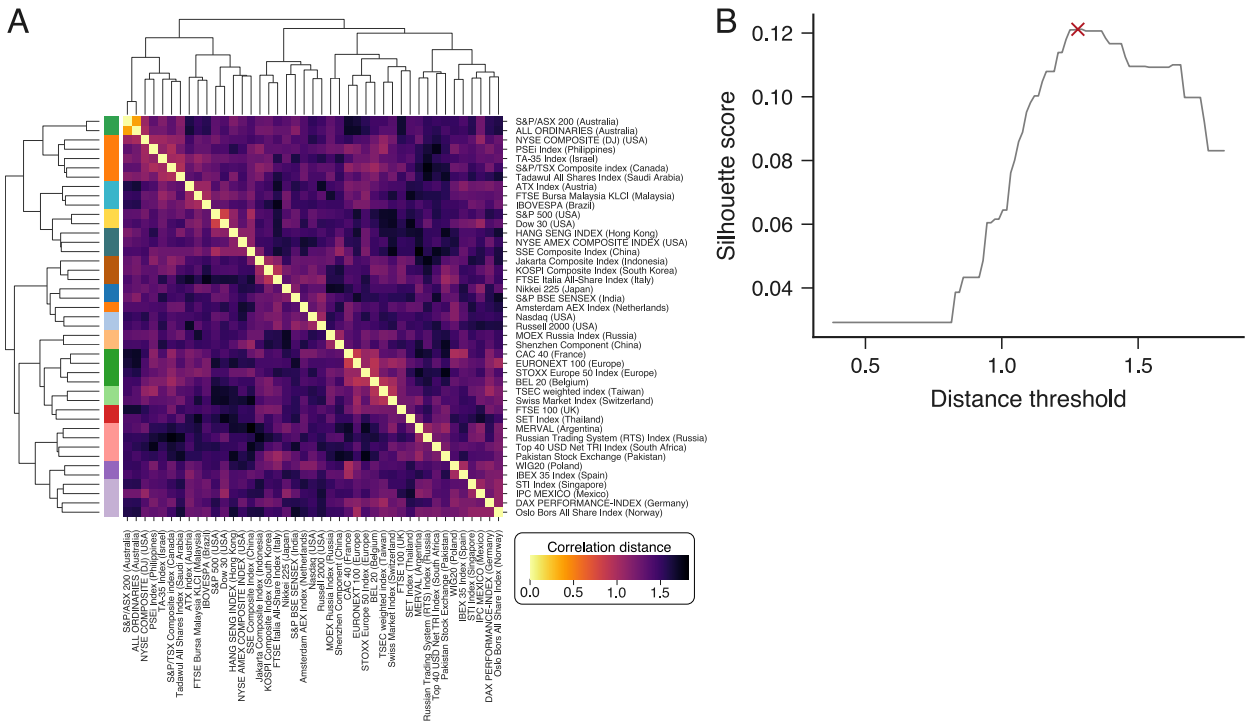

Figura 4.7: Organização hierárquica de longo prazo dos padrões de eficiência dos principais mercados de ações mundiais. (A) Matriz representando a distância de correlação entre todos os pares de séries temporais da entropia dos mercados de ações. Os dendrogramas mostram os resultados dos agrupamentos hierárquicos baseados no método da miníma variância de Ward. (B) Coeficiente de silhueta a partir dos grupos obtidos ao segmentar o dendrograma em diferentes limiares de distância. Os retângulos coloridos localizados abaixo dos ramos dos dendrogramas no painel (A) indicam os 16 grupos obtidos ao segmentar o dendrograma na distância limiar que maximiza o valor do coeficiente de silhueta (indicada pela cruz vermelha).

Em seguida, utilizamos o método da miníma variância de Ward [118] (já utilizado na seção 2.6) para construir uma representação em dendrograma da matriz de distâncias, que também é mostrada na figura 4.7A. Nossos resultados indicam que os mercados de ações formam uma estrutura hierárquica de acordo com a evolução da eficiência informacional no longo prazo. No entanto, não observamos estruturas de agrupamento grandes na matriz de distâncias. Na verdade, ao determinar o número de grupos maximizando o coeficiente de silhueta [119] (mesmo procedimento que foi adotado nas seções 2.6 e 4.2.2), como mostra a figura 4.7B, temos 16 grupos dos quais 15 consistem em apenas alguns mercados (o maior grupo possui 5 mercados) e 1 grupo contém um único mercado. A figura 4.8 mostra que esses grupos de mercados exibem perfil temporal de $H(t)$ semelhantes no longo prazo, porém essa análise global não captura movimentos de curto prazo de $H(t)$ entre os mercados de ações.

Para investigar o comportamento coletivo de curto prazo da eficiência dos mercados de ações, amostramos a série temporal de $H(t)$ com uma janela móvel de um ano e criamos um *ranking* de eficiência dos mercados baseado no valor médio de $H(t)$ dentro de cada janela temporal. Em seguida, investigamos a estabilidade desses *rankings* de eficiência estimando o coeficiente de correlação de *rankings* de Kendall [190] (Kendall-τ) entre todos os pares possíveis de janelas temporais. Essa análise resulta na matriz de correlação mostrada na figura 4.9A, na qual linhas e colunas representam a última data de cada janela. Por definição, os elementos da diagonal dessa matriz são unitários (ou seja, o *ranking* de eficiência em uma dada janela é perfeitamente correlacionado consigo mesmo). Os valores em uma dada linha ou coluna indicam o quão similar é o *ranking* das eficiências daquela data com *rankings* em datas passadas e futuras. Assim, devemos obter grandes estruturas em bloco diagonais com valores altos do coeficiente de Kendall-τ se esses *rankings* de eficiência forem estáveis por períodos longos. No entanto, observamos pequenos blocos diagonais com aproximadamente um mês de largura, indicando que os *rankings* de eficiência são estáveis durante períodos curtos.

Além disso, calculamos a distância de correlação da eficiência $H(t)$ entre todos os pares de mercados para cada janela temporal e aplicamos o mesmo procedimento de agrupamento usado para toda a série temporal de $H(t)$ (ou seja, criamos o dendrograma a partir do método da miníma variância de Ward, segmentado no valor máximo do coeficiente de silhueta). Essa abordagem produz grupos de mercados com evolução similar da eficiência em cada janela temporal. Comparamos a estabilidade temporal desses grupos estimando o índice de Rand ajustado [191] entre todos os resultados de agrupamentos em cada janela temporal. Esse coeficiente mede a concordância entre dois agrupamentos ao contar os pares de elementos atribuídos ao mesmo grupo, controlando para a sobreposição esperada ao acaso. Valores desse índice próximos a zero indicam que os dois agrupamentos não são mais similares do que partições aleatórias, enquanto valores iguais a 1 indicam concordância perfeita entre esses

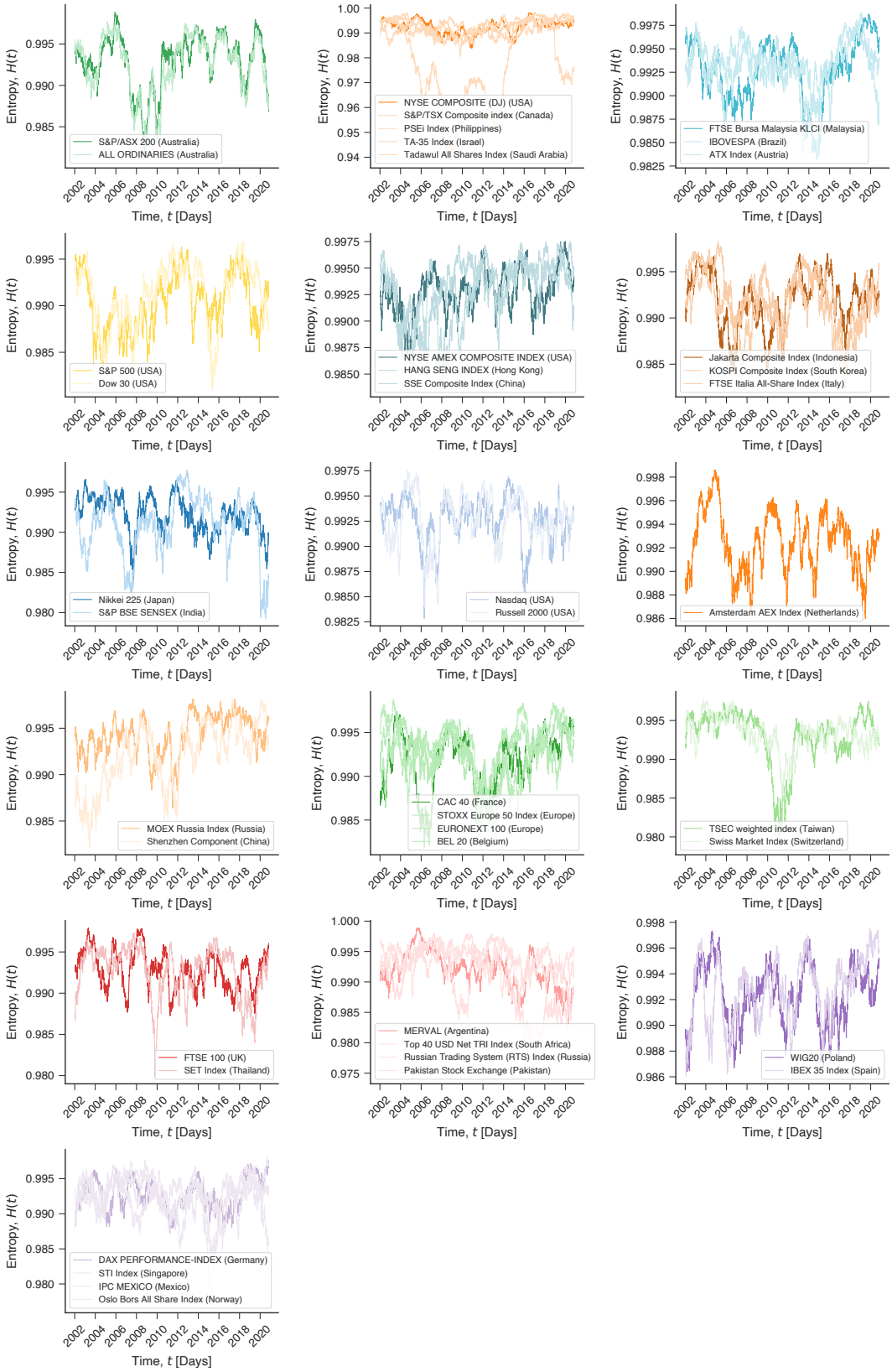

Figura 4.8: Evolução temporal da entropia de permutação $H(t)$ para os 43 mercados de ações agrupados de acordo com os grupos obtidos a partir da dinâmica de longo prazo de $H(t)$.

agrupamentos. A figura 4.9B mostra a matriz dos índices de Rand ajustados para todos os resultados de pares de grupos. As pequenas estruturas em forma de blocos na diagonal dessa matriz indicam que grupos de mercados com perfil similar de $H(t)$ permanecem estáveis por aproximadamente 4 meses.

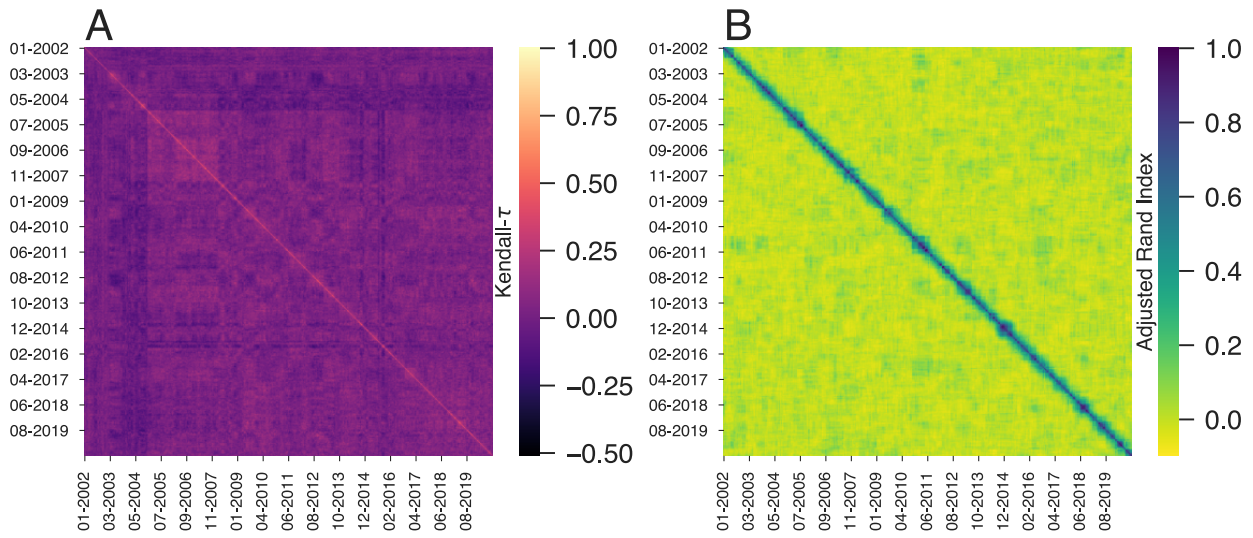

Figura 4.9: Estabilidade de curto prazo dos *rankings* de eficiência e dos agrupamentos dos mercados de ações. (A) Matriz representando o coeficiente de correlação tau de Kendall (Kendall-τ) entre todos os pares de *rankings* de eficiência dos mercados de ações calculado dentro de uma janela móvel de 1 ano. Notamos a formação de pequenos blocos diagonais com largura de 1 a 2 meses, indicando que os *rankings* de eficiência mudam com o passar do tempo. (B) Gráfico de matriz do índice de Rand ajustado que estima a concordância entre agrupamentos de mercados com perfis temporais de eficiência similares em diferentes janelas de tempo. Também observamos a existência de pequenos blocos diagonais com largura de aproximadamente 4 meses, mostrando que grupos com perfis de eficiência similares não permanecem estáveis durante longos períodos.

Os resultados da figura 4.9 demonstram que os padrões coletivos de curto prazo na evolução da eficiência dos mercados de ações mudam com o tempo e são diferentes daqueles obtidos para escalas temporais longas (figura 4.7A). Os padrões dinâmicos que encontramos indicam que partições simples não são suficientes para capturar as interações complexas entre mercados financeiros e motivam o uso de uma abordagem diferente que considere essas interações emaranhadas. Para isso, propomos a construção de uma rede complexa na qual os nós representam mercados de ações, e as conexões indicam dois mercados que são agrupados juntos pelo menos uma vez ao longo do tempo. Também assumimos que a conexão entre dois mercados de ações é ponderada pelo número de vezes que esses mercados específicos são agrupados juntos. Essa representação permite agregar a informação de curto prazo em uma imagem global, na qual mercados de ações cujas dinâmicas de eficiência estejam correlacionadas durante algum período apareçam conectados. Além disso, os pesos dessas conexões indicam a intensidade das interações entre os mercados.

A figura 4.10A mostra essa representação de rede para os 43 mercados do nosso estudo. Observamos que essa rede complexa forma um grafo completo já que apresenta todas as possíveis conexões entre todos os mercados de ações. Portanto, os mercados de ações mundiais são fortemente globalizados quanto à sua eficiência, de modo que períodos simultâneos de alta ou baixa eficiência geralmente envolvem um grande número de mercados. Esse resultado sugere a existência de risco sistêmico para o "espalhamento" de estados de baixa eficiência; mas, ao mesmo tempo, indica que estados de alta eficiência também podem emergir globalmente. Embora a densidade dessa rede financeira seja máxima, as interações entre os mercados de ações não são uniformemente distribuídas. O coeficiente de Gini dos pesos das conexões é 0,18 (em uma escala em que zero representa igualdade perfeita e um máxima desigualdade), sugerindo que alguns mercados podem ter um impacto maior na dinâmica da eficiência do sistema. A figura 4.10B mostra o *ranking* de centralidade baseado no PageRank [192], no qual os índices *Amsterdam AEX Index* (Países Baixos) e o *KOSPI Composite Index* (Coréia do Sul) emergem como os mercados mais influentes, enquanto os dois índices russos (*MOEX Russia Index* e *Russian Trading System (RTS) Index*) são os menos influentes para a dinâmica da eficiência dos mercados de ações mundial.

A desigualdade na distribuição dos pesos das conexões também sugere que a rede financeira da figura 4.10A pode ter uma estrutura modular na qual grupos de mercados são mais similares entre si do que com outros grupos. Para investigar essa possível estrutura modular, usamos a abordagem conhecida por modelo estocástico de blocos (SBM) [193–195]. Esse método possui a vantagem de estimar diretamente as probabilidades marginais de que a rede seja particionada em um certo número de grupos e também a probabilidade de um nó pertencer a um grupo específico durante o processo de inferência. Testamos diferentes modelos estocástico de blocos (SBM) para ajustar os dados da nossa rede: SBM usual, SBM com correção de grau (DCSBM), SBM aninhado e DCSBM aninhado. Consideramos o melhor modelo como aquele que possui o menor comprimento mínimo de descrição [195], com base na evidência estatística de que o modelo não está confundindo flutuações estocásticas com a estrutura modular real. Isso significa que os grupos nessa rede não podem ter surgido a partir de flutuações estocásticas, como ocorre em grafos totalmente aleatórios [196]. Encontramos que o modelo estocástico de blocos aninhado sem correção de grau (SBM aninhado, veja tabela 4.2) é o que melhor descreve nossa rede.

Esse modelo resulta em dois módulos representados pelas cores dos nós na figura 4.10A. Coletamos ainda as partições para 10.000 varreduras do algoritmo de aceitação-rejeição Metropolis-Hastings via método de cadeia de Markov Monte Carlo [197] com vários movimentos para amostrar partições da rede, em intervalos de 10 varreduras. Usando essas partições, estimamos a probabilidade marginal do número de módulos da rede, conforme mostra a figura 4.11. Esse resultado reforça que a estrutura modular com dois módulos é a mais provável para nossa rede financeira. Também estimamos as probabilidades marginais

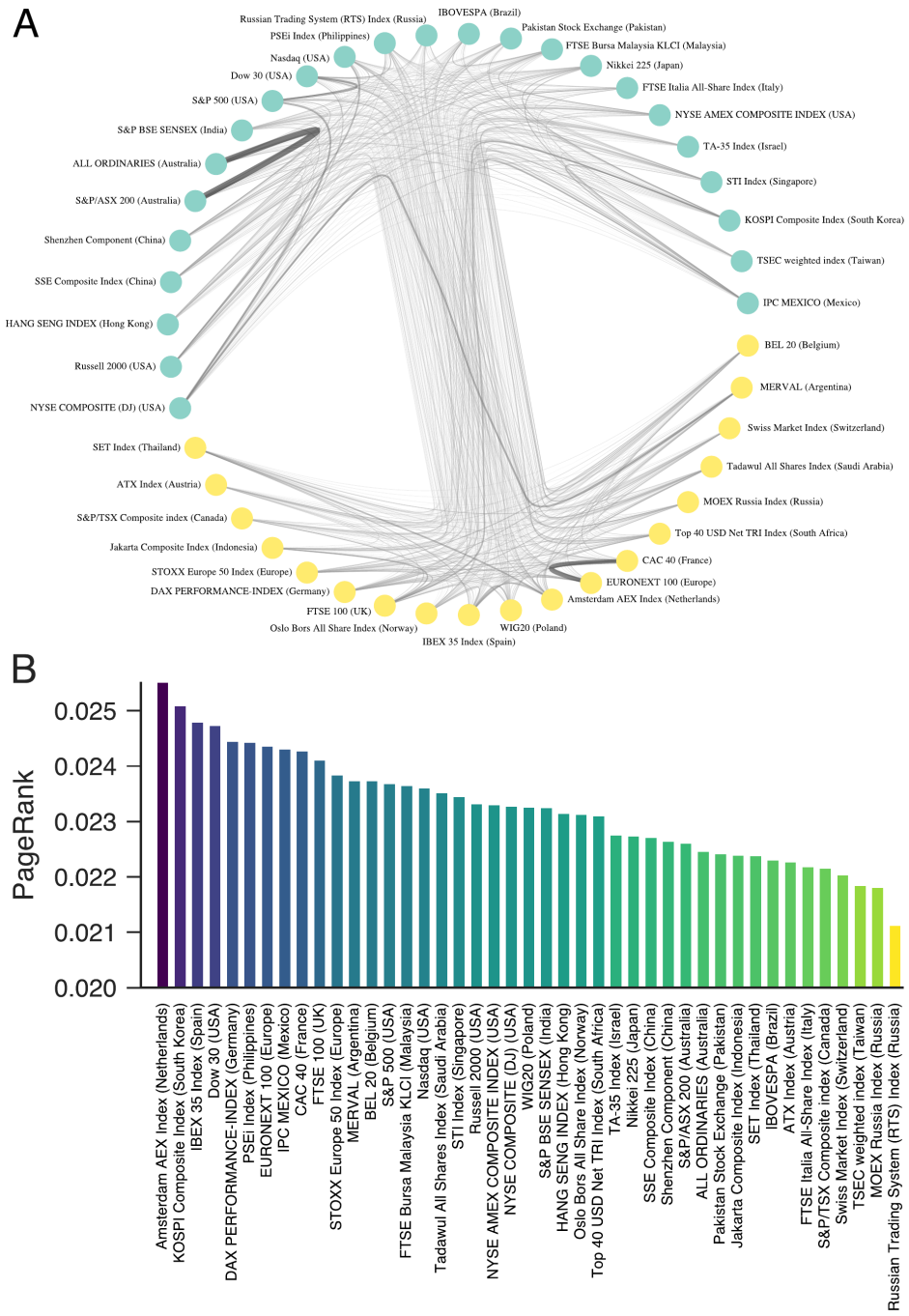

Figura 4.10: Rede financeira dos mercados de ações exibindo tendências semelhantes de eficiência no curto prazo. (A) Os nós dessa rede representam os mercados de ações e as conexões são feitas entre mercados que aparecem pelo menos uma vez no mesmo grupo com relação à similaridade de curto prazo na evolução da eficiência. Essas conexões são ponderadas pelo número de vezes que pares de mercados aparecem no mesmo grupo. Usando o modelo estocástico de blocos aninhado, identificamos dois módulos na rede indicados pelas cores diferentes dos nós. Nessa visualização, as larguras das conexões são proporcionais aos pesos. (B) *Ranking* da centralidade baseada no PageRank indicando os mercados de ações mais influentes para a dinâmica da eficiência dos índices dos mercados de ações.

Modelo	Abreviação	Comprimento mínimo de descrição
Modelo estocástico de blocos	SBM	55
Modelo estocástico de blocos aninhado com correção de grau	DCSBM aninhado	48
Modelo estocástico de blocos com correção de grau	DCSBM	41
Modelo estocástico de blocos aninhado	SBM aninhado	29

Tabela 4.2: Seleção do modelo baseada no comprimento mínimo de descrição.

de um nó pertencer a um grupo em nossa rede financeira. Os resultados indicam que a vasta maioria dos mercados são quase sempre atribuídos à mesma partição. Todos os modelos de blocos estocásticos foram implementados com a biblioteca *graph-tool* [198] na linguagem Python.

O maior módulo dessa rede (com 24 índices) compreende mercados dos Estados Unidos (6 índices), países da Ásia-Pacífico (14 índices) e 4 outros mercados do Brasil, Itália, Israel e México. O segundo maior módulo inclui os outros 19 índices: 12 da Europa e outros mercados da Argentina, Canadá, Indonésia, Rússia, Arábia Saudita, África do Sul e Tailândia. Apesar da existência de muitas exceções, a distância geográfica parece desempenhar algum papel nessas partições. Porém, mais importante do que entender as particularidades de cada módulo, é a emergência dessa estrutura modular. Embora as associações entre os mercados de ações sejam bastante emaranhadas, essa estrutura modular sugere que alguns grupos de mercados são mais similares entre si, de modo que estados de baixa ou de alta eficiência têm maior probabilidade de ocorrer dentro de cada um dos módulos.

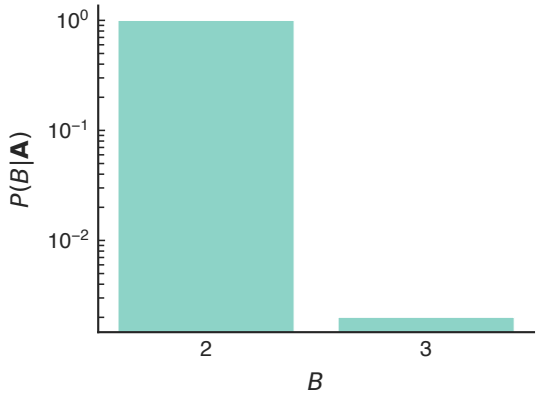

Figura 4.11: Probabilidade marginal das partições $P(B|\mathbf{A})$ em nossa rede financeira. O gráfico de barras mostra a função densidade de probabilidade $P(B|\mathbf{A})$ de que a rede $\mathbf{A}$ seja particionada em B módulos. A distribuição está concentrada em $B = 2$, indicando que dois módulos representam a estrutura modular mais provável para nossa rede.

4.4 Conclusão

Nesse capítulo, apresentamos dois estudos em grande escala sobre a eficiência informacional dos mercados de criptomoedas e de ações. No caso do mercado de criptomoedas, utilizamos a entropia e complexidade de permutação calculadas em janelas móveis sobre as séries temporais do retorno logarítmico dos preços. Nossa pesquisa mostrou que o mercado de criptomoedas é, em grande parte, compatível com a hipótese do mercado eficiente, com somente 20% das criptomoedas sendo menos que 20% do tempo informacionalmente eficientes. Por outro lado, mais da metade do mercado de criptomoedas é mais de 60% do tempo informacionalmente eficiente e 37% das 437 criptomoedas em nosso estudo são mais de 80% do tempo informacionalmente eficientes, resultado que está de acordo com pesquisas anteriores na mesma linha [172,173], nas quais foram reportados resultados similares considerando apenas a moeda Bitcoin.

Além disso, com base em uma análise da eficiência informacional ao longo do tempo e após agrupar as criptomoedas que possuem perfis temporais similares, identificamos as moedas digitais que seguem os mesmos altos e baixos com respeito à eficiência informacional. Ao passo que as particularidades da composição dos diferentes grupos podem ser de interesse para investidores que procuram um portfólio diversificado, mostramos que dentro dos quatro grupos identificados, as moedas mais jovens parecem seguir a tendência das moedas pioneiras. As similaridades nas tendências temporais entre as moedas mais jovens e as pioneiras nos leva a conjecturar que o mercado de criptomoedas pode se tornar tão aderente à hipótese de mercado eficiente quanto os mercados de ações.

Na análise sobre mercado de ações, apresentamos uma investigação dos padrões de eficiência dinâmica de 43 principais mercados de ações durante os últimos 20 anos. Para isso, também usamos uma abordagem inspirada em Física, na qual a eficiência de um mercado de ações em um determinado momento é definida como a entropia de permutação dentro de janelas móveis dos retornos logarítmicos dos índices dos mercados. Nossos resultados indicam que os mercados de ações podem ser hierarquicamente organizados em grupos de mercados que possuem tendências semelhantes de eficiência de longo prazo. No entanto, também descobrimos que esses grupos de longo prazo não são suficientes para compreender a dinâmica coletiva da eficiência dos mercados. Na verdade, nossa pesquisa revelou que padrões coletivos de curto prazo na evolução da eficiência variam com o tempo e são diferentes daqueles de longo prazo. Também observamos que os *rankings* de eficiência dos mercados de ações só são estáveis em períodos relativamente curtos, não mais do que um ou dois meses. Similarmente, o agrupamento de mercados com perfis de eficiência semelhantes é estável somente por aproximadamente quatro meses.

Por causa desses fatos, propusemos uma representação na forma de rede complexa na qual os nós são os mercados e as conexões entre eles indicam mercados que aparecem no

mesmo grupo pelo menos uma vez durante os 20 anos de dados que possuímos. Além disso, consideramos que as conexões entre um par de mercados são ponderadas pelo número de vezes que esse par aparece no mesmo grupo com relação à dinâmica da eficiência de curto prazo. Nossos resultados mostram que essa rede financeira é completamente conectada, indicando que a eficiência dos mercados de ações é fortemente emaranhada e globalizada. Trabalhos anteriores já haviam mostrado que falhas sistêmicas nos mercados de ações emergem a partir de um processo de sincronização [183] que ocorre em redes sociais [85, 199] e podem causar bolhas financeiras que eventualmente estouram [185]. Outros estudos também sugerem a existência de correlações fortes entre estratégias de fundos de investimentos como uma causa de risco sistêmico e propagação de choque [189]. Nesse contexto, a rede financeira de nosso trabalho sugere um risco sistêmico de propagação de estados de baixa eficiência mas, ao mesmo tempo, indica que estados de alta eficiência também podem emergir em um nível global. Embora nossa rede financeira forme um grafo completo, os pesos das conexões entre os mercados de ações não são uniformemente distribuídos, permitindo identificar os mercados mais influentes. Além do mais, constatamos que essa rede financeira possui uma estrutura modular que compreende dois grupos de mercados cuja dinâmica de eficiência é mais similar dentro dos grupos do que com mercados fora dos grupos. Portanto, apesar dos mercados de ações serem bastante emaranhados em termos dos seus perfis de eficiência, a estrutura modular indica que estados de baixa eficiência e estados de alta eficiência são mais prováveis de emergir dentro desses grupos.

Visto que a eficiência nos mercados de ações pode ser uma oportunidade de lucro, bem como um sinal precoce de uma crise financeira iminente, seria interessante incorporar nossa abordagem em modelos que tentam prever os preços dos mercados de ações para medir os riscos de transações financeiras. Por exemplo, pesquisadores usaram o *Google Trends* para informar investidores quando comprar ou vender ações e descobriram que sua estratégia era 326% melhor do que a estratégia tradicional *buy-and-hold* [184]. Apesar da impressionante melhora dessa estratégia, ela não informa os volumes de compra ou de venda, ou os riscos envolvidos nas transações propostas. Assim, quantificar a dependência no tempo da eficiência das ações pode ajudar agentes econômicos a quantificar os riscos de suas transações.

Com isso, esperamos que nossas abordagens inspiradas na Física para determinar a eficiência dos mercado de criptomoedas e de ações possa levar a novas sinergias frutíferas entre a Física e a Economia [7], e que os resultados apresentados contribuam para o amplo campo de pesquisa sobre os mercados financeiros e os recém desenvolvidos mercados de criptomoedas.

Conclusões e Perspectivas

Nesse trabalho, apresentamos uma combinação de métodos de Física Estatística e de Ciência de Dados para estudar quatro sistemas complexos bastante distintos [32–36]. Essa combinação mostrou-se muito propícia para investigar padrões e comportamentos coletivos desses sistemas. Essa sinergia ocorre porque, fundamentalmente, essas disciplinas compartilham do mesmo objetivo: descrever como elementos de conjuntos de dados ou sistemas interagem e influenciam uns aos outros, gerando tendências e características emergentes.

É interessante mencionar que, em quatro das cinco análises apresentadas, apenas duas medidas de complexidade estatística foram empregadas como quantificadores do sistema: a entropia de permutação e a complexidade estatística de permutação. Esse fato destaca bem a versatilidade e o potencial dessa abordagem baseada em conceitos de Física Estatística. Combinadas a essas medidas, associamos técnicas de aprendizagem de máquina visando extrair ainda mais informações a respeito do comportamento dos sistemas que estudamos.

No capítulo 2, dentre outros resultados, quantificamos aspectos e conceitos qualitativos ensinados por historiadores da arte por meio da análise de padrões espaciais de imagens de obras de arte de uma grande base de dados. Nessa mesma linha, no capítulo 3, analisamos imagens digitais e propusemos duas abordagens para extrair propriedades físicas de cristais líquidos a partir de imagens de texturas desses materiais. A primeira utiliza as duas medidas de complexidade estatística combinadas com métodos de aprendizagem de máquina. Já a segunda, emprega redes convolucionais neurais diretamente nas texturas de cristal líquido, dispensando a extração manual de características das imagens. Ambas abordagens alcançam altas precisões em tarefas de previsão de propriedades físicas de cristais líquidos. Além disso, no capítulo 4, quantificamos diversos padrões relacionados à eficiência informacional de mercados de criptomoedas e de ações por meio da análise de séries financeiras via entropia e complexidade estatística de permutação. Entre outros resultados, mostramos que o recém criado mercado de criptomoedas parece estar evoluindo para um estado de maior eficiência

informacional e que a eficiência no mercado de ações é um fenômeno coletivo. Destacamos que todas nossas abordagens foram, relativamente, bem-sucedidas em seus objetivos. Esse fato indica a utilidade prática dessa associação entre a Física de Sistemas Complexos e a Ciência de Dados.

Por fim, como perspectiva de estudos futuros podemos citar, o estudo de artistas específicos e da distribuição e uso das cores nas obras de arte para quantificar como essas características estéticas podem ter evoluído entre diferentes pintores e estilos. Uma outra possibilidade, envolvendo texturas de cristais líquidos, é investigar outros materiais mais complexos e outras propriedades físicas. Quanto aos mercados financeiros, devido a recente introdução do mercado de criptomoedas, nossa pesquisa colabora para que possamos ter uma compreensão melhor de seu comportamento e compará-lo com a atividade do mercado financeiro tradicional. Nesse sentido, nosso trabalho representa uma pequena contribuição na tentativa de preencher essas lacunas e contribuir para o avanço das aplicações interdisciplinares da Física.

APÊNDICE A

Obtenção das texturas de cristais líquidos

Os experimentos e simulações descritos neste apêndice são oriundos de uma colaboração com pesquisadores do grupo de cristais líquidos da UTFPR do câmpus Apucarana.

A.1 Simulações pelo método de Monte Carlo para gerar texturas nemáticas

Para obter as texturas nemáticas analisadas nas seções 3.2.1, 3.3.1 e 3.3.2, simulamos um sistema composto de *headless spins* localizados sobre os pontos de uma rede cúbica tridimensional de dimensões $N_x \times N_y \times N_z$ (com $N_x = N_y = 100$ e $N_z = 20$). Esses *spins* possuem direção representadas por vetores unitários $\vec{u}_i$ [$i = (1, 2, \ldots, N)$, com $N = N_x N_y N_z = 200.000$]. Os *spins* da primeira camada na direção z são fixos e apontam na direção y, enquanto aqueles da última camada são fixos ao longo da direção x. Essas duas camadas de *spins* fixos imitam uma região de superfície (denotada por $\mathcal{S}$) e fornecem uma direção de ancoragem que muda o alinhamento dos *spins* ao longo da amostra. Os outros *spins* na região *bulk* (denotado por $\mathcal{B}$) interagem com os vizinhos mais próximos via potencial de Lebwohl-Lasher [200] com condições de contorno periódicas ao longo das direções x e y. A hamiltoniana desse sistema pode ser escrita como

$$U_{\mathrm{N}} = \frac{1}{2} \sum_{\substack{i,j \in \mathcal{B} \\ i \neq j}} \Phi_{ij} + J \sum_{\substack{i \in \mathcal{B} \\ j \in \mathcal{S}}} \Phi_{ij}, \tag{A.1}$$

na qual J é magnitude da energia de ancoragem e

$$\Phi_{ij} = -\epsilon_{ij}\left(\frac{3}{2}\cos(\vec{u}_i \cdot \vec{u}_j) - \frac{1}{2}\right), \tag{A.2}$$

com $\epsilon_{ij} = \epsilon$ quando i e j são vizinhos mais próximos e zero caso contrário.

Todas as texturas foram obtidas com $J = 1$ e os *spins* do *bulk* estão inicialmente alinhados formando um ângulo com respeito à direção x [$u_i = (\cos(0{,}3), \sin(0{,}3), 0)$] para evitar o aparecimento de defeitos instáveis [201, 202]. As atualizações da rede ocorrem por meio do algoritmo de Metropolis [203]. Cada nova configuração é gerada seguindo a técnica de Barker-Watts [204] e é aceita com probabilidade $\exp(-\frac{\Delta U}{k_B T})$, na qual ΔU é a diferença de energia entre os estados antigo e novo, T a temperatura e k_B a constante de Boltzmann. Um passo de Monte Carlo é completado quando, em média, todos os *spins* são atualizados.

As simulações começam com uma dada temperatura reduzida $T_r = k_B T/\epsilon$ e o sistema é inicialmente simulado por 10^4 passos de Monte Carlo para evitar comportamentos transientes. Em seguida, consideramos mais 10^4 passos de Monte Carlo para estimar o parâmetro de ordem médio em cada camada do sistema. O parâmetro de ordem local é o maior autovalor da matriz de ordem

$$Q_{ab} = \left\langle u_i^{(a)} u_i^{(b)} - \delta_{ab} \right\rangle, \tag{A.3}$$

na qual $u_i^{(a)}$ e $u_i^{(b)}$ são as componentes a e b do vetor unitário $\vec{u}_i$ associado ao o i-ésimo *spin* e δ_{ab} representa a delta de Kronecker. O parâmetro de ordem da amostra (p) é definido como o valor médio de Q_{ab} sobre cada camada livre do sistema. Esse modelo é conhecido por passar por uma transição nemática-isotrópica em $T_c = 1{,}1232$ [205] que diminui para $T_c = 1{,}1075$ ao considerar uma amostra confinada com condições de contorno híbridas (como no nosso caso). Para simulações com temperaturas reduzidas próximas a essa temperatura crítica, o sistema é inicialmente simulado por 9×10^4 passos de Monte Carlo para evitar comportamentos transientes relacionados à transição de fase. As texturas são obtidas utilizando as médias dos últimos 50 passos de Monte Carlo via metodologia de Stokes-Muller [206]. Esse procedimento consiste em tratar a luz incidente (paralela a direção z) como um vetor de Stokes e descrever cada ponto como uma matriz de Muller. Consideramos $n_e = 1{,}66$ para o índice de refração extraordinário, $n_0 = 1{,}50$ para o índice de refração ordinário, espessura da amostra de 5,3 μm e comprimento de onda de 545 nm para a luz incidente. A textura resultante é representada por uma matriz de dimensões 100×100, cujos elementos representam a intensidade de luz transmitida em torno de um ponto particular na área da superfície.

A.2 Procedimento experimental para obtenção das texturas do cristal líquido E7

As texturas experimentais analisadas nas seções 3.2.2 e 3.3.4 foram obtidas via microscopia óptica de luz polarizada utilizando amostras de um cristal líquido em diferentes temperaturas. A amostra consiste em capilares retangulares sem tratamento de superfície (300 μm $\times$ 4 mm) contendo a mistura E7 na temperatura de 70 °C para evitar alinhamento de fluxo. Em seguida, as amostras são resfriadas a temperatura ambiente e colocadas em um controlador de temperatura acoplado ao arranjo do microscópio óptico de luz polarizado. Iniciamos os procedimentos experimentais tirando fotografias das texturas das amostras em 40 °C. As amostras são lentamente aquecidas a uma taxa constante de 0,2 °C por minuto e fotografias são produzidas a cada 90 segundos até atingir a temperatura de 55 °C. Para temperaturas maiores, a taxa de aquecimento é reduzida a 0,05 °C por minuto e as fotos são produzidas a cada 60 segundos até atingir a temperatura de 61 °C.

Todas os arquivos das imagens obtidas estão no formato PNG com dimensões de 2047 *pixels* de largura por 1532 *pixels* de altura e 24 bits por *pixel*, sendo 8 *bits* para cada uma das três cores no espaço de cores RGB. Isso significa que cada *pixel* da imagem é caracterizado por uma entre 256 intensidades de vermelho (R), verde (G) e azul (B), permitindo $256^3 = 16.777.216$ variações de cores. Esses arquivos podem ser representados por uma matriz de três camadas com dimensões n_x (a largura da imagem) por n_y (a altura da imagem), cujas camadas correspondem a cada uma das cores do espaço RGB e cujos elementos representam as intensidades das cores (variando de 0 a 255). Essa representação é análoga a utilizada para o estudo com as obras de arte (capítulo 2). Calculamos $0{,}2125R + 0{,}7154G + 0{,}0721B$, ou seja, uma média ponderada sobre as três camadas, sendo que R, G e B representam as intensidades das cores vermelho, verde e azul de cada *pixel*, respectivamente. Esse procedimento corresponde a transformação em escala de cinza chamada luminância (ou refletância) [108], conforme também discutimos no capítulo 2. Ao final do processo, temos uma única matriz para cada arquivo de imagem, a partir da qual os valores de H e C são calculados ou essa matriz é usada diretamente como dado de entrada nas redes neurais convolucionais.

A.3 Simulações das texturas colestéricas via teoria elástica contínua

As texturas colestéricas investigadas nas seções 3.2.3 e 3.3.3 foram obtidas por meio da teoria elástica contínua. Em particular, usamos a abordagem de Landau-de Gennes [149] para descrever a densidade de energia F associada às variações no parâmetro de ordem

tensorial Q em torno do estado de equilíbrio. Se empregarmos x_1, x_2, e x_3 para representar as coordenadas espaciais, a densidade de energia pode ser escrita como

$$\begin{aligned} F = & \frac{L_1}{2}\frac{\partial Q_{ij}}{\partial x_k}\frac{\partial Q_{ij}}{\partial x_k} + \frac{L_2}{2}\frac{\partial Q_{ij}}{\partial x_j}\frac{\partial Q_{ik}}{\partial x_k} \\ & + \frac{L_3}{2}Q_{ij}\frac{\partial Q_{kl}}{\partial x_i}\frac{\partial Q_{kl}}{\partial x_j} + \frac{4\pi}{\eta}L_q\epsilon_{ikl}Q_{ij}\frac{\partial Q_{lj}}{\partial x_k} \\ & + \frac{A}{2}Q_{ij}Q_{ji} + \frac{B}{3}Q_{ij}Q_{jk}Q_{ki} \\ & + \frac{C}{4}Q_{ij}Q_{jk}Q_{kl}Q_{li}\,, \end{aligned} \tag{A.4}$$

na qual L_1, L_2, L_3, e L_q são constantes elásticas, A, B, e C são parâmetros termodinâmicos e η é o comprimento do passo colestérico. Aqui, assumimos soma implícita nos índices repetidos (notação de Einstein). A evolução temporal das componentes de Q_{ij} é dada por

$$\Gamma\frac{\partial Q_{ij}}{\partial t} = \left(\frac{\partial F(Q)}{\partial Q_{ij}} - \frac{d}{dx_k}\frac{\partial F(Q)}{\partial Q_{ij,k}}\right)\,, \tag{A.5}$$

na qual Γ é a viscosidade rotacional do cristal líquido, t é o tempo e $Q_{ij,k}$ é a derivada de Q_{ij} relativa a x_k.

O sistema de equações A.5 é resolvido numericamente via método das diferenças finitas em uma grade uniforme contendo $200 \times 200 \times 20$ pontos. Todas as unidades de distância são reescaladas pelas distâncias da grade $\delta x_1 = 1$ nm, $\delta x_2 = 1$ nm e $\delta x_3 = 1$ nm. Os parâmetros utilizados para o cristal líquido são $A = -0{,}348$ MJ/(Km^3), $B = -2{,}133$ MJ/m^3, $C = 1{,}733$ MJ/m^3, $\Gamma = 0{,}3$ Pa s, $L_1 = 2{,}6$ pN, $L_2 = 2{,}6$ pN, $L_3 = 0{,}76$ pN, e $L_q = 1{,}86$ pN. Esses parâmetros foram obtidos da literatura e são conhecidos por descrever aproximadamente o cristal líquido 5CB [149]. Consideramos diferentes valores para o passo η e o método de integração temporal de Dormand-Prince de quarta ordem [207]. A condição inicial é escolhida aleatoriamente a partir de uma distribuição uniforme e condições de contorno periódicas são consideradas em todas as direções para evitar efeitos de superfície. Finalmente, as texturas ópticas resultantes são geradas aplicando o método 2×2 de Jones [148] com as células do cristal líquido colocadas entre polarizadores cruzados. Essa abordagem é a mesma utilizada na referência [208] para gerar imagens ópticas.

Referências Bibliográficas

[1] Lohr, S. The origins of "big data": An etymological detective story. http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/ (2013). Acessado em 12/01/2022.

[2] 2.5 quintillion bytes of data generated everyday – Top data science trends 2020. https://us.sganalytics.com/blog/2-5-quintillion-bytes-of-data-generated-everyday-top-data-science-trends-2020/ (2020). Acessado em 12/01/2022.

[3] Big data and what it means. https://www.uschamberfoundation.org/bhq/big-data-and-what-it-means. Acessado em 12/01/2022.

[4] Data, data everywhere. http://www.economist.com/special-report/2010/02/25/data-data-everywhere (2010). Acessado em 12/01/2022.

[5] LinkedIn's 2020 emerging jobs report. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf (2020). Acessado em 12/01/2022.

[6] Relatório LinkedIn das profissões emergentes 2020. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_Brazil.pdf (2020). Acessado em 12/01/2022.

[7] Mantegna, R. N. & Stanley, H. E. *Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 1999).

[8] Sumpter, D. J. *Collective Animal Behavior* (Princeton University Press, 2010).

[9] Faghmous, J. H. & Kumar, V. A big data guide to understanding climate change: The case for theory-guided data science. *Big Data* **2**, 155–163 (2014).

[10] Galam, S. *Sociophysics: A Physicist's Modeling of Psycho-Political Phenomena* (Springer Science & Business Media, Berlin, 2012).

[11] Dawood, F. S. *et al.* Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a H1N1 virus circulation: A modelling study. *The Lancet Infectious Diseases* **12**, 687–695 (2012).

[12] Liu, Q.-H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y. & Vespignani, A. Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences* **115**, 12680–12685 (2018).

[13] Ponte, C., Carmona, H. A., Oliveira, E. A., Caminha, C., Lima, A. S., Andrade, J. S. & Furtado, V. Tracing contacts to evaluate the transmission of COVID-19 from highly exposed individuals in public transportation. *Scientific Reports* **11**, 1–11 (2021).

[14] Hino, M., Benami, E. & Brooks, N. Machine learning for environmental monitoring. *Nature Sustainability* **1**, 583–588 (2018).

[15] Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C. & Barabási, A.-L. Quantifying reputation and success in art. *Science* **362**, 825–829 (2018).

[16] Kosack, S., Coscia, M., Smith, E., Albrecht, K., Barabási, A.-L. & Hausmann, R. Functional structures of US state governments. *Proceedings of the National Academy of Sciences* **115**, 11748–11753 (2018).

[17] Gerlach, M., Farb, B., Revelle, W. & Amaral, L. A. N. A robust data-driven approach identifies four personality types across four large data sets. *Nature Human Behaviour* **2**, 735 (2018).

[18] Picoli, S., Antonio, F. J., Itami, A. S. & Mendes, R. S. Power-law relaxation in human violent conflicts. *The European Physical Journal B* **90**, 1–5 (2017).

[19] Ribeiro, H. V., Alves, L. G. A., Martins, A. F., Lenzi, E. K. & Perc, M. The dynamical structure of political corruption networks. *Journal of Complex Networks* **6**, 989–1003 (2018).

[20] Alves, L. G. A., Ribeiro, H. V. & Rodrigues, F. A. Crime prediction through urban metrics and statistical learning. *Physica A* **505**, 435–443 (2018).

[21] Vieira, D. S., Picoli, S. & Mendes, R. S. Robustness of sentence length measures in written texts. *Physica A* **506**, 749–754 (2018).

[22] Alves, L. G. A., Andrade, J. S., Hanley, Q. S. & Ribeiro, H. V. The hidden traits of endemic illiteracy in cities. *Physica A* **515**, 566–574 (2019).

[23] Ribeiro, H. V., Rybski, D. & Kropp, J. P. Effects of changing population or density on urban carbon dioxide emissions. *Nature Communications* **10**, 1–9 (2019).

[24] Vieira, D. S., Riveros, J. M. E., Jauregui, M. & Mendes, R. S. Anomalous diffusion behavior in parliamentary presence. *Physical Review E* **99**, 042141 (2019).

[25] Ribeiro, H. V., Sunahara, A. S., Sutton, J., Perc, M. & Hanley, Q. S. City size and the spreading of COVID-19 in Brazil. *PLoS One* **15**, e0239699 (2020).

[26] Sunahara, A. S., Perc, M. & Ribeiro, H. V. Association between productivity and journal impact across disciplines and career age. *Physical Review Research* **3**, 033158 (2021).

[27] Pessa, A. A. B., Zola, R. S., Perc, M. & Ribeiro, H. V. Determining liquid crystal properties with ordinal networks and machine learning. *Chaos, Solitons & Fractals* **154**, 111607 (2022).

[28] Bandt, C. & Pompe, B. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters* **88**, 174102 (2002).

[29] López-Ruiz, R., Mancini, H. L. & Calbet, X. A statistical measure of complexity. *Physics Letters A* **209**, 321–326 (1995).

[30] Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A. & Fuentes, M. A. Distinguishing noise from chaos. *Physical Review Letters* **99**, 154102 (2007).

[31] James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* (Springer, New York, 2013).

[32] Sigaki, H. Y. D., Perc, M. & Ribeiro, H. V. History of art paintings through the lens of entropy and complexity. *Proceedings of the National Academy of Sciences* **115**, E8585–E8594 (2018).

[33] Sigaki, H. Y. D., de Souza, R. F., de Souza, R. T., Zola, R. S. & Ribeiro, H. V. Estimating physical properties from liquid crystals textures via machine learning and complexity-entropy methods. *Physical Review E* **99**, 013311 (2019).

[34] Sigaki, H. Y. D., Lenzi, E. K., Zola, R. S., Perc, M. & Ribeiro, H. V. Learning physical properties of liquid crystals with deep convolutional neural networks. *Scientific Reports* **10**, 7664 (2020).

[35] Sigaki, H. Y. D., Perc, M. & Ribeiro, H. V. Clustering patterns in efficiency and coming-of-age of the cryptocurrency market. *Scientific Reports* **9**, 1–9 (2019).

[36] Alves, L. G. A., Sigaki, H. Y. D., Perc, M. & Ribeiro, H. V. Collective dynamics of stock market efficiency. *Scientific Reports* **10**, 1–10 (2020).

[37] Ribeiro, H. V., Zunino, L., Lenzi, E. K., Santoro, P. A. & Mendes, R. S. Complexity-entropy causality plane as a complexity measure for two-dimensional patterns. *PLOS ONE* **7**, e40689 (2012).

[38] Bissell, M. Reproducibility: The risks of the replication drive. *Nature News* **503**, 333 (2013).

[39] Anônimo. Reality check on reproducibility. *Nature News* **533**, 437 (2016).

[40] Ward, A., Baldwin, T. O. & Antin, P. B. Research data: Silver lining to irreproducibility. *Nature* **532**, 177–177 (2016).

[41] Anônimo. Repetitive flaws. *Nature News* **529**, 256 (2016).

[42] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 (1948).

[43] Martin, M. T., Plastino, A. & Rosso, O. A. Statistical complexity and disequilibrium. *Physics Letters A* **311**, 126–132 (2003).

[44] Kowalski, A. M., Martín, M. T., Plastino, A., Rosso, O. A. & Casas, M. Distances in probability space and the statistical complexity setup. *Entropy* **13**, 1055–1075 (2011).

[45] Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52**, 479–487 (1988).

[46] Ribeiro, H. V., Jauregui, M., Zunino, L. & Lenzi, E. K. Characterizing time series via complexity-entropy curves. *Physical Review E* **95**, 062106 (2017).

[47] Rényi, A. On measures of entropy and information (The Regents of the University of California, 1961).

[48] Jauregui, M., Zunino, L., Lenzi, E. K., Mendes, R. S. & Ribeiro, H. V. Characterization of time series via Rényi complexity–entropy curves. *Physica A* **498**, 74–85 (2018).

[49] Wootters, W. K. Statistical distance and Hilbert space. *Physical Review D* **23**, 357–362 (1981).

[50] Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J. & Stanley, H. E. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E* **65**, 041905 (2002).

[51] Martin, M. T., Plastino, A. & Rosso, O. A. Generalized statistical complexity measures: geometrical and analytical properties. *Physica A* **369**, 439 (2006).

[52] McCarthy, J. What is Artificial Intelligence? *Stanford University* (2007).

[53] Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**, 210–229 (1959).

[54] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386 (1958).

[55] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[56] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015).

[57] Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (The MIT Press, 2016).

[58] Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L. & Webster, D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* **2**, 158 (2018).

[59] Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* **23**, 89–109 (2001).

[60] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).

[61] Kim, K., Kim, S., Lee, Y. H., Lee, S. H., Lee, H. S. & Kim, S. Performance of the deep convolutional neural network based magnetic resonance image scoring algorithm for differentiating between tuberculous and pyogenic spondylitis. *Scientific Reports* **8**, 13124 (2018).

[62] Lindsey, R. *et al.* Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences* **115**, 11591–11596 (2018).

[63] Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G. & Seuntjens, J. Deep learning in head & neck cancer outcome prediction. *Scientific Reports* **9**, 2764 (2019).

[64] Lee, C. S., Tyring, A. J., Wu, Y., Xiao, S., Rokem, A. S., DeRuyter, N. P., Zhang, Q., Tufail, A., Wang, R. K. & Lee, A. Y. Generating retinal flow maps from structural optical coherence tomography with artificial intelligence. *Scientific Reports* **9**, 5694 (2019).

[65] Abdeltawab, H. *et al.* A novel CNN-based CAD system for early assessment of transplanted kidney dysfunction. *Scientific Reports* **9**, 5948 (2019).

[66] Kim, K.-J. Financial time series forecasting using support vector machines. *Neurocomputing* **55**, 307–319 (2003).

[67] Cambria, E. & White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* **9**, 48–57 (2014).

[68] Hartley, R. I. & Zisserman, A. *Multiple View Geometry in Computer Vision* (Cambridge University Press, 2000).

[69] Szeliski, R. *Computer Vision: Algorithms and Applications* (Springer, 2010).

[70] Zhang, Y. & Pang, J. Distance and friendship: A distance-based model for link prediction in social networks. In *Asia-Pacific Web Conference*, 55–66 (Springer, 2015).

[71] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics (Springer, New York, 2013).

[72] Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**, 175–185 (1992).

[73] Oliphant, T. E. *A guide to NumPy* (Trelgol Publishing USA, 2006).

[74] Jones, E., Oliphant, T., Peterson, P. *et al.* SciPy: Open source scientific tools for Python (2001).

[75] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[76] Prince, S. J. D. *Computer Vision: Models, Learning, and Inference* (Cambridge University Press, 2012).

[77] Chollet, F. *Deep Learning with Python* (Manning Publications, Greenwich, 2017).

[78] Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

[79] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012).

[80] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034 (2015).

[81] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

[82] Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL https://www.tensorflow.org/.

[83] Chollet, F. *et al.* Keras. https://keras.io (2015).

[84] Wang, Z., Bauch, C. T., Bhattacharyya, S., d'Onofrio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M. & Zhao, D. Statistical physics of vaccination. *Physics Reports* **664**, 1–113 (2016).

[85] Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S. & Szolnoki, A. Statistical physics of human cooperation. *Physics Reports* **687**, 1–51 (2017).

[86] Birkhoff, G. D. *Aesthetic Measure* (Harvard University Press Cambridge, Cambridge, 1933).

[87] Taylor, R. P., Micolich, A. P. & Jonas, D. Fractal analysis of Pollock's drip paintings. *Nature* **399**, 422 (1999).

[88] Jones-Smith, K. & Mathur, H. Fractal analysis: Revisiting Pollock's drip paintings. *Nature* **444**, E9–E10 (2006).

[89] Taylor, R. P., Micolich, A. P. & Jonas, D. Fractal analysis: Revisiting Pollock's drip paintings (reply). *Nature* **444**, E10 (2006).

[90] Taylor, R. P., Guzman, R., Martin, T. P., Hall, G. D. R., Micolich, A. P., Jonas, D., Scannell, B. C., Fairbanks, M. S. & Marlow, C. A. Authenticating Pollock paintings using fractal geometry. *Pattern Recognition Letters* **28**, 695–702 (2007).

[91] Jones-Smith, K., Mathur, H. & Krauss, L. M. Drip paintings and fractal analysis. *Physical Review E* **79**, 046111 (2009).

[92] De la Calleja, E. M., Cervantes, F. & De la Calleja, J. Order-fractal transitions in abstract paintings. *Annals of Physics* **371**, 313–322 (2016).

[93] Boon, J. P., Casti, J. & Taylor, R. P. Artistic forms and complexity. *Nonlinear Dynamics-Psychology and Life Sciences* **15**, 265 (2011).

[94] Alvarez-Ramirez, J., Ibarra-Valdez, C. & Rodriguez, E. Fractal analysis of Jackson Pollock's painting evolution. *Chaos, Solitons & Fractals* **83**, 97–104 (2016).

[95] Pedram, P. & Jafari, G. R. Mona Lisa: The stochastic view and fractality in color space. *International Journal of Modern Physics C* **19**, 855–866 (2008).

[96] Taylor, R. Pollock, Mondrian and the nature: Recent scientific investigations. *Chaos and Complexity Letters* **1**, 29 (2004).

[97] Hughes, J. M., Graham, D. J. & Rockmore, D. N. Quantification of artistic style through sparse coding analysis in the drawings of pieter bruegel the elder. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1279–1283 (2010).

[98] Shamir, L. Computer analysis reveals similarities between the artistic styles of Van Gogh and Pollock. *Leonardo* **45**, 149–154 (2012).

[99] Elsa, M. & Zenit, R. Topological invariants can be used to quantify complexity in abstract paintings. *Knowledge-Based Systems* **126**, 48–55 (2017).

[100] Castrejon-Pita, J. R., Castrejón-Pita, A. A., Sarmiento-Galán, A. & Castrejón-Garcıa, R. Nasca lines: A mystery wrapped in an enigma. *Chaos* **13**, 836–838 (2003).

[101] Koch, M., Denzler, J. & Redies, C. $1/f^2$ characteristics and isotropy in the Fourier power spectra of visual art, cartoons, comics, mangas, and different categories of photographs. *PLOS ONE* **5**, e12268 (2010).

[102] Montagner, C., Linhares, J. M. M., Vilarigues, M. & Nascimento, S. M. C. Statistics of colors in paintings and natural scenes. *Jounral of the Optical Society of America A* **33**, A170–A177 (2016).

[103] Stork, D. G. & Coddington, J. (eds.). *Computer image analysis in the study of art, Proceedings of SPIE*, vol. 6810 (2008).

[104] Stork, D. G., Coddington, J. & Bentkowska-Kafel, A. (eds.). *Computer vision and image analysis of art, Proceedings of SPIE*, vol. 7531 (2010).

[105] Stork, D. G., Coddington, J. & Bentkowska-Kafel, A. (eds.). *Computer vision and image analysis of art II, Proceedings of SPIE*, vol. 7869 (2011).

[106] Kim, D., Son, S.-W. & Jeong, H. Large-scale quantitative analysis of painting arts. *Scientific Reports* **4**, 7370 (2014).

[107] Lee, B., Kim, D., Sun, S., Jeong, H. & Park, J. Heterogeneity in chromatic distance in images and characterization of massive painting data set. *PLOS ONE* **13**, e0204430 (2018).

[108] van der Walt, S., Schönberger, J. L., Nunez Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T. & the scikit-image contributors. scikit-image: Image processing in Python. *PeerJ* **2**, e453 (2014).

[109] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).

[110] Wölfflin, H. *Principles of Art History: The Problem of the Development of Style in Later Art* (Dover, Mineola, 1950).

[111] Riegl, A. *Historical Grammar of the Visual Arts* (Zone Book, New York, 2004).

[112] Gaiger, J. The analysis of pictorial style. *The British Journal of Aesthetics* **42**, 20–36 (2002).

[113] Danto, A. C. & Goehr, L. *After the End of Art: Contemporary Art and the Pale of History* (Princeton University Press, Princeton, 1997).

[114] Kleiner, F. S. *Gardner's Art Through the Ages: The Western Perspective* (Wadsworth Publishing, Boston, 2013).

[115] Hodge, A. N. *A History of Art: Painting from Giotto to the Present Day* (Arcturus Publishing Limited, London, 2013).

[116] Blatt, S. J. & Blatt, E. S. *Continuity and Change in Art: The Development of Modes of Representation* (Routledge, New York, 1984).

[117] Dunn, O. J. Multiple Comparisons among means. *Journal of the American Statistical Association* **56**, 52–64 (1961).

[118] Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236–244 (1963).

[119] Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

[120] Chowdhury, G. G. *Introduction to Modern Information Retrieval* (Facet publishing, London, 2010).

[121] Müller, A. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O'Reilly, Sebastopol, 2016).

[122] Zujovic, J., Gandy, L., Friedman, S., Pardo, B. & Pappas, T. N. Classifying paintings by artistic genre: An analysis of features & classifiers. In *IEEE International Workshop on Multimedia Signal Processing*, 1–5 (2009).

[123] Agarwal, S., Karnick, H., Pant, N. & Patel, U. Genre and style based painting classification. In *IEEE Winter Conference on Applications of Computer Vision*, 588–594 (2015).

[124] Chen, F., Brown, G. M. & Song, M. Overview of 3-d shape measurement using optical methods. *Optical Engineering* **39**, 10–23 (2000).

[125] Wang, Z., Li, H., Zhu, Y. & Xu, T. Review of plant identification based on image processing. *Archives of Computational Methods in Engineering* **24**, 637–654 (2017).

[126] Prost, J. & de Gennes, P. G. *The Physics of Liquid Crystals* (Oxford University Press, 1995).

[127] Zola, R. S., Evangelista, L., Yang, Y.-C. & Yang, D.-K. Surface induced phase separation and pattern formation at the isotropic interface in chiral nematic liquid crystals. *Physical Review Letters* **110**, 057801 (2013).

[128] Zheng, Z.-g., Zola, R. S., Bisoyi, H. K., Wang, L., Li, Y., Bunning, T. J. & Li, Q. Controllable dynamic zigzag pattern formation in a soft helical superstructure. *Advanced Materials* **29**, 1701903 (2017).

[129] Nemati, H., Yang, D.-K., Cheng, K.-L., Liang, C.-C., Shiu, J.-W., Tsai, C.-C. & Zola, R. Effect of surface alignment layer and polymer network on the helfrich deformation in cholesteric liquid crystals. *Journal of Applied Physics* **112**, 124513 (2012).

[130] Arsenault, L.-F., Lopez-Bezanilla, A., von Lilienfeld, O. A. & Millis, A. J. Machine learning for many-body physics: The case of the Anderson impurity model. *Physical Review B* **90**, 155136 (2014).

[131] Kusne, A. G. *et al.* On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Scientific Reports* **4**, 6367 (2014).

[132] Cubuk, E. D., Schoenholz, S. S., Rieser, J. M., Malone, B. D., Rottler, J., Durian, D. J., Kaxiras, E. & Liu, A. J. Identifying structural flow defects in disordered solids using machine-learning methods. *Physical Review Letters* **114**, 108001 (2015).

[133] Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big–deep–smart data in imaging for guiding materials design. *Nature Materials* **14**, 973 (2015).

[134] Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nature Physics* **13**, 431 (2017).

[135] Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications* **5**, 4308 (2014).

[136] Mukund, N., Abraham, S., Kandhasamy, S., Mitra, S. & Philip, N. S. Transient classification in LIGO data using difference boosting neural network. *Physical Review D* **95**, 104059 (2017).

[137] Dreissigacker, C., Sharma, R., Messenger, C., Zhao, R. & Prix, R. Deep-learning continuous gravitational waves. *Physical Review D* **100**, 044009 (2019).

[138] Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O. & Baker, N. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv:1706.06689* (2017).

[139] Ma, W., Cheng, F. & Liu, Y. Deep-learning-enabled on-demand design of chiral metamaterials. *ACS Nano* **12**, 6326–6334 (2018).

[140] Zhang, Z., Schott, J. A., Liu, M., Chen, H., Lu, X., Sumpter, B. G., Fu, J. & Dai, S. Prediction of carbon dioxide adsorption via deep learning. *Angewandte Chemie International Edition* **58**, 259–263 (2019).

[141] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547 (2018).

[142] Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C. & Agrawal, A. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports* **8**, 17593 (2018).

[143] Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nature Communications* **9**, 2775 (2018).

[144] Wei, J., Chu, X., Sun, X.-Y., Xu, K., Deng, H.-X., Chen, J., Wei, Z. & Lei, M. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).

[145] Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. & Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95 (2019).

[146] Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **4**, 053208 (2016).

[147] Schwab, K. The Fourth Industrial Revolution. *Foreign Affairs* (2015).

[148] Wu, S. & Yang, D. *Fundamentals of Liquid Crystal Devices*. Wiley Series in Display Technology (Wiley, Chichester, 2006).

[149] Ravnik, M. & Žumer, S. Landau - de Gennes modelling of nematic liquid crystal colloids. *Liquid Crystals* **36**, 1201–1214 (2009).

[150] Smith, L. N. & Topin, N. Deep convolutional neural network design patterns. *arXiv preprint arXiv:1611.00847* (2016).

[151] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[152] Cootner, P. *The Random Character of Stock Market Prices* (MIT Press, Cambridge, 1964).

[153] Malkiel, B. G. & Fama, E. F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* **25**, 383–417 (1970).

[154] Fama, E. F. The behavior of stock-market prices. *The Journal of Business* **38**, 34–105 (1965).

[155] Malkiel, B. G. The efficient market hypothesis and its critics. *Journal of Economic Perspectives* **17**, 59–82 (2003).

[156] Preis, T. & Stanley, H. E. Bubble trouble. *Physics World* **24**, 29 (2011).

[157] Sornette, D. *Why Stock Markets Crash: Critical Events in Complex Financial Systems* (Princeton University Press, Princeton, 2017).

[158] Fox, J. & Sklar, A. *The Myth of the Rational Market: A History of Risk, Reward, and Delusion on Wall Street* (Harper Business, New York, 2009).

[159] Zunino, L., Tabak, B. M., Pérez, D. G., Garavaglia, M. & Rosso, O. A. Inefficiency in Latin-American market indices. *The European Physical Journal B* **60**, 111–121 (2007).

[160] Zunino, L., Tabak, B. M., Figliola, A., Pérez, D., Garavaglia, M. & Rosso, O. A multifractal approach for stock market inefficiency. *Physica A* **387**, 6558–6566 (2008).

[161] Zunino, L., Zanin, M., Tabak, B. M., Pérez, D. G. & Rosso, O. A. Forbidden patterns, permutation entropy and stock market inefficiency. *Physica A* **388**, 2854–2864 (2009).

[162] Zunino, L., Zanin, M., Tabak, B. M., Pérez, D. G. & Rosso, O. A. Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. *Physica A* **389**, 1891–1901 (2010).

[163] Zunino, L., Bariviera, A. F., Guercio, M. B., Martinez, L. B. & Rosso, O. A. On the efficiency of sovereign bond markets. *Physica A* **391**, 4342–4349 (2012).

[164] Szarek, D., Łukasz Bielak & Wyłomańska, A. Long-term prediction of the metals' prices using non-Gaussian time-inhomogeneou stochastic process. *Physica A* **555**, 124659 (2020).

[165] Wang, L. & Li, L. Long-range correlation and predictability of Chinese stock prices. *Physica A* **549**, 124384 (2020).

[166] Filho, T. M. R. & Rocha, P. M. Evidence of inefficiency of the Brazilian stock market: The IBOVESPA future contracts. *Physica A* **543**, 123200 (2020).

[167] Sánchez-Granero, M., Balladares, K., Ramos-Requena, J. & Trinidad-Segovia, J. Testing the efficient market hypothesis in Latin American stock markets. *Physica A* **540**, 123082 (2020).

[168] Urquhart, A. The inefficiency of Bitcoin. *Economics Letters* **148**, 80–82 (2016).

[169] Bariviera, A. F., Basgall, M. J., Hasperué, W. & Naiouf, M. Some stylized facts of the Bitcoin market. *Physica A* **484**, 82–90 (2017).

[170] Zhang, W., Wang, P., Li, X. & Shen, D. The inefficiency of cryptocurrency and its cross-correlation with Dow Jones industrial average. *Physica A* **510**, 658–670 (2018).

[171] Bariviera, A. F. The inefficiency of Bitcoin revisited: A dynamic approach. *Economics Letters* **161**, 1–4 (2017).

[172] Nadarajah, S. & Chu, J. On the inefficiency of Bitcoin. *Economics Letters* **150**, 6–9 (2017).

[173] Tiwari, A. K., Jana, R., Das, D. & Roubaud, D. Informational efficiency of Bitcoin – An extension. *Economics Letters* **163**, 106–109 (2018).

[174] Bariviera, A. F., Zunino, L. & Rosso, O. A. An analysis of high-frequency cryptocurrencies prices dynamics using permutation-information-theory quantifiers. *Chaos* **28**, 075511 (2018).

[175] Alvarez-Ramirez, J., Rodriguez, E. & Ibarra-Valdez, C. Long-range correlations and asymmetry in the bitcoin market. *Physica A* **492**, 948–955 (2018).

[176] Dimitrova, V., Fernández-Martínez, M., Sánchez-Granero, M. & Trinidad Segovia, J. Some comments on bitcoin market (in)efficiency. *PLOS ONE* **14**, e0219243 (2019).

[177] Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43–49 (1978).

[178] Sakoe, H. & Chiba, S. Dynamic-programming approach to continuous speech recognition. In *Proceedings of the International Congress of Acoustics, Budapest* (1971).

[179] Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y. Time-series clustering – A decade review. *Information Systems* **53**, 16–38 (2015).

[180] Stanley, H. E., Amaral, L. A., Gabaix, X., Gopikrishnan, P. & Plerou, V. Similarities and differences between physics and economics. *Physica A* **299**, 1–15 (2001).

[181] Gopikrishnan, P., Plerou, V., Amaral, L. A. N., Meyer, M. & Stanley, H. E. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E* **60**, 5305 (1999).

[182] Mantegna, R. N. & Stanley, H. E. Scaling behaviour in the dynamics of an economic index. *Nature* **376**, 46–49 (1995).

[183] Johansen, A., Ledoit, O. & Sornette, D. Crashes as critical points. *International Journal of Theoretical and Applied Finance* **3**, 219–255 (2000).

[184] Preis, T., Moat, H. S. & Stanley, H. E. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* **3**, 1684 (2013).

[185] Sornette, D. Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth. *Proceedings of the National Academy of Sciences* **99**, 2522–2529 (2002).

[186] Aroussi, R. *et al.* Yahoo! Finance market data downloader (2020). Disponível em: https://github.com/ranaroussi/yfinance.

[187] Journal, W. S. Wall Street Journal market data (2020). Disponível em: https://www.wsj.com/market-data.

[188] del Canto, A. B. investpy - Financial Data Extraction from Investing.com with Python (2020). Available at: https://github.com/alvarobartt/investpy.

[189] Delpini, D., Battiston, S., Caldarelli, G. & Riccaboni, M. Systemic risk from investment similarities. *PLOS ONE* **14**, e0217141 (2019).

[190] Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).

[191] Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850 (1971).

[192] Brin, S. & Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks* **56**, 3825–3833 (2012).

[193] Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks* **5**, 109–137 (1983).

[194] Funke, T. & Becker, T. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE* **14**, e0215296 (2019).

[195] Peixoto, T. P. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* **4**, 011047 (2014).

[196] Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* **106**, 22073–22078 (2009).

[197] Peixoto, T. P. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E* **89**, 012804 (2014).

[198] Peixoto, T. P. The graph-tool Python library. *figshare* (2014).

[199] Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81**, 591 (2009).

[200] Lebwohl, P. A. & Lasher, G. Nematic-liquid-crystal order – A Monte Carlo calculation. *Physical Review A* **6**, 426 (1972).

[201] Chiccoli, C., Evangelista, L., Pasini, P., Teixeira de Souza, R. & Zannoni, C. Effect of surface anchoring on creation of defects in a nematic film: A Monte Carlo simulation. *Molecular Crystals and Liquid Crystals* **614**, 137–143 (2015).

[202] Chiccoli, C., Evangelista, L., Omori, E., Pasini, P., Teixeira-Souza, R. & Zannoni, C. Computer simulation of a nematic hybrid cell: The effects of elastic anisotropy. *Molecular Crystals and Liquid Crystals* **649**, 86–93 (2017).

[203] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

[204] Barker, J. & Watts, R. Structure of water; A Monte Carlo calculation. *Chemical Physics Letters* **3**, 144–145 (1969).

[205] Fabbri, U. & Zannoni, C. A Monte Carlo investigation of the Lebwohl-Lasher lattice model in the vicinity of its orientational phase transition. *Molecular Physics* **58**, 763–788 (1986).

[206] Berggren, E., Zannoni, C., Chiccoli, C., Pasini, P. & Semeria, F. Computer simulations of nematic droplets with bipolar boundary conditions. *Physical Review E* **50**, 2929 (1994).

[207] Dormand, J. R. & Prince, P. J. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics* **6**, 19–26 (1980).

[208] Seč, D., Porenta, T., Ravnik, M. & Žumer, S. Geometrical frustration of chiral ordering in cholesteric droplets. *Soft Matter* **8**, 11982 (2012).