UNIVERSIDADE ESTADUAL DE MARINGÁ DEPARTAMENTO DE FÍSICA

HIGOR YUDI DUENHA SIGAKI

FÍSICA ESTATÍSTICA APLICADA AO ESTUDO DE OBRAS DE ARTE

Maringá, Agosto de 2017.

UNIVERSIDADE ESTADUAL DE MARINGÁ DEPARTAMENTO DE FÍSICA

HIGOR YUDI DUENHA SIGAKI

FÍSICA ESTATÍSTICA APLICADA AO ESTUDO DE OBRAS DE ARTE

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Física.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, Agosto de 2017.

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

Sigaki, Higor Yudi Duenha

S574f

Física estatística aplicada ao estudo de obras de arte / Higor Yudi Duenha Sigaki. -- Maringá, 2017. 57 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro. Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física, 2017.

1. Análise de dados - Física estatística. 2. Física estatística. 3. Sistemas complexos - Física estatística. 4. Entropia. 5. Pintura - Análise de dados. 6. Artes - Análise 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. 530.1595

GVS-003775

Agradecimentos

Agradeço em primeiro lugar, a Deus por minha vida, família e amigos. Ele que me presenteou com saúde, força, persistência e determinação para superar as atribulações da vida e seguir adiante com esse trabalho.

À minha mãe Maria, ao meu pai Hélio e aos meus irmãos, Débora, Eduardo e Hugo por, apesar dos contratempos, sempre me incentivarem, cada um a sua maneira.

À minha namorada Larissa, por ser compreensiva e paciente o bastante ao entender as demandas necessárias para realizar esse trabalho.

Agradeço imensamente ao Prof. Dr. Haroldo Valentin Ribeiro, pela oportunidade, paciência e disposição ao me orientar durante o mestrado. Um profissional ímpar, o qual tive a sorte de trabalhar junto durante esse período.

A todos os colegas da pós-graduação, que são de extrema importância, principalmente por proporcionarem momentos valiosos de descontração.

A esta universidade e a todos os professores do Departamento de Física, que foram fundamentais para minha formação.

Resumo

Neste trabalho, focamos nossa atenção em quantificar aspectos das artes visuais utilizando ferramentas da Física Estatística. Em particular, utilizamos duas medidas de complexidade estatística: a entropia e a complexidade estatística de permutação. Aplicamos essas duas medidas de complexidade às imagens obtidas de uma das maiores bases de dados disponível online sobre artes visuais, a Wiki Art. Para lidar com essa grande quantidade de dados (mais de cem mil imagens) e todos os metadados associados (como data, estilo e gênero), utilizamos pacotes e programas escritos em linguagem Python. Empregando o plano complexidade-entropia, mostramos que é possível distinguir dentre os vários estilos artísticos dessas obras de arte. Aplicamos um método de agrupamento hierárquico visando encontrar grupos de estilos que são (em algum sentido) similares uns aos outros. Além disso, verificamos uma evolução temporal clara das medidas de complexidade com o passar do tempo, a qual tentamos associar fatos da história da arte. Argumentamos, ainda, que essas características evolutivas podem ser associadas com mudanças nos estilos artísticos predominantes de cada período. De maneira geral, acreditamos que nosso estudo representa uma contribuição inicial para a aplicação de ferramentas e métodos da Física Estatística na investigação quantitativa das artes visuais e sua história.

Palavras-chave: Análise de dados. Física Estatística. Complexidade. Entropia. Pinturas. Artes.

Abstract

We focus our attention on quantifying aspects of visual arts by using tools from statistical physics. In particular, we use two statistical complexity measures: the permutation entropy and the statistical complexity. We apply these two complexity measures to the images obtained from one of the largest online available databases about visual arts, the Wiki Art. In order to deal with this huge amount of data (more than one hundred thousand images) and all associated metadata (such as date, style and genre), we employ packages and programs written in Python language. By using the complexity-entropy plane, we show that it is possible to distinguish among the several artistic styles. We also apply a hierarchical clustering method aiming to find groups of styles which are (in some sense) similar to each other. Moreover, we verify a clear evolutive trend of the complexity measures over time, a feature that we try to associate with some facts of art history. We further argue, that this evolutive feature can be associated with changes in the predominant styles of each particular period of time. Our study thus represents an initial contribution for applying tools and methods from statistical physics to the quantitative investigation of visual arts and its history.

Keywords: Data analysis. Statistical physics. Complexity. Entropy. Paintings. Art.

Sumário

In	Introdução			
1	Ent	ropia e complexidade de permutação	9	
	1.1	A entropia de permutação e a abordagem de Bandt e Pompe	9	
	1.2	Outras medidas de complexidade estatística	16	
	1.3	Plano complexidade-entropia	19	
	1.4	Generalização do plano complexidade-entropia para sistemas multidimensionais	22	
2	A base de dados sobre artes visuais $Wiki\ Art$			
	2.1	Descrição e apresentação dos dados	27	
	2.2	Utilizando a linguagem de programação <i>Python</i> e seus pacotes para extração		
		e processamento dos dados	30	
3	Plano complexidade-entropia aplicado às imagens de obras de arte			
	3.1	Independência das medidas de complexidade estatística com a dimensão das		
		imagens	37	
	3.2	Distinguindo os estilos artísticos no plano complexidade-entropia	39	
	3.3	Agrupando os estilos artísticos utilizando um método de agrupamento hierár-		
		quico	42	
	3.4	Evolução temporal das medidas de complexidade estatística no plano complexidad	de-	
		entropia	49	
C	onclu	ısões e perspectivas	52	
\mathbf{R}_{0}	e <mark>fer</mark> ê	ncias bibliográficas	53	

Introdução

Os sistemas que são o foco de estudo dessa dissertação e de nosso grupo de pesquisa são conhecidos por sistemas complexos [1–10]. No sentido mais intuitivo, complexo se contrapõe ao que é considerado simples ou fácil. Porém, definir o que é complexo, no âmbito da Física e da Matemática, não é tão trivial quanto parece. Como o leitor irá notar, esse trabalho é baseado em duas medidas de complexidade estatística: a entropia de permutação [11] e a complexidade estatística de permutação [12,13]. Logo no primeiro capítulo, apresentamos, na seção 1.2, uma tentativa de explicar e ilustrar o que, dentro do contexto dessas medidas, entendemos por complexidade. Ao estudar sistemas complexos, em geral, tentamos entender como um sistema composto de muitas partes interagentes produz comportamentos coletivos emergentes. Nesse trabalho, estudamos obras de arte por meio dessa perspectiva, empregando ferramentas para análise de dados e conceitos da Física Estatística.

De modo geral, uma obra de arte é composta de muitos elementos, como cores, formas, geometrias e técnicas de pincelada, os quais se relacionam de maneira muito complexa e praticamente impossível de se quantificar por uma expressão matemática. Desse ponto de vista, uma obra de arte pode ser considerada um sistema complexo cujo efeito coletivo, quando apresentada a um apreciador, é estimular um conjunto de emoções, pensamentos e outras reações sensoriais únicas. É muito pouco provável que dois indíviduos tenham as mesmas sensações e impressões quando expostos a uma mesma obra de arte.

As artes visuais têm um papel fundamental no desenvolvimento de nossa sociedade. Elas são utilizadas como forma de comunicação para expressar ideias e emoções muito antes do desenvolvimento da escrita. A evolução da pintura, por exemplo, reflete enormemente o contexto histórico-social em que a sociedade está imersa. Muitos movimentos artísticos emergiram ao representar os fatos e sentimentos vividos pela sociedade da época. Desse modo, entender como as artes visuais evoluíram ao longo dos séculos é de extrema importância para entender um pouco mais sobre o desenvolvimento de nossa cultura e, consequentemente, de

nossa história.

Por ser uma forma de expressão criativa muito subjetiva, existe uma dificuldade inerente em quantificar aspectos das artes visuais. Nesse sentido, utilizando apenas as duas medidas de complexidade estatística mencionadas anteriormente, abordamos esse problema na tentativa de responder perguntas como:

- Será que é possível classificar estilos artísticos utilizando essas medidas de complexidade estatística?
- Essas quantidades quantificam a evolução desses movimentos artísticos ao longo do tempo?
- Será que podemos encontrar grupos de estilos que tenham algum grau de similaridade?

Acreditamos que a resposta para essas perguntas é afirmativa e, para respondê-las, construímos uma grande base de dados de imagens digitais de obras de arte (quase 140 mil imagens) a partir da *internet*, na qual aplicamos essas duas medidas de complexidade. Com essa abordagem, mostramos que é possível contornar o caráter subjetivo intrínseco das artes visuais e estudá-las de maneira quantitativa.

Sendo assim, esse trabalho encontra-se dividido em três capítulos. No capítulo 1, apresentamos as técnicas e as medidas de complexidade que foram alvo de nosso estudo. No capítulo seguinte, mostramos como os dados utilizados para as análises foram obtidos e processados. Por fim, no capítulo 3, apresentamos as análises dos dados e nossos resultados.

Entropia e complexidade de permutação

Neste capítulo, apresentamos as técnicas que foram utilizadas ao longo do desenvolvimento desse trabalho visando identificar padrões na dinâmica de movimentos artísticos por meio da análise de obras de arte. Especificamente, estudamos duas medidas de complexidade estatística conhecidas como entropia de permutação e complexidade estatística de permutação [11–14]. Notamos que o termo "permutação" diz respeito ao método de simbolização proposto por Bandt e Pompe [11] para obter a distribuição de probabilidade relacionada aos estados acessíveis do sistema em questão, em nosso caso, as imagens das obras de arte.

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

O método proposto por Bandt e Pompe [11] visa, fundamentalmente, encontrar uma medida de complexidade "natural" para uma dada série temporal. O problema, segundo argumentam, é que existem várias medidas e significados diferentes de complexidade para comparar essas séries e distingui-las de acordo com suas diversas características, por exemplo, entre regular ou caótica. Entretanto, a maioria delas não pode ser aplicada para uma série temporal arbitrária, ou seja, cada tipo de série demanda um algoritmo específico para realizar a medida em questão. Essa, por sua vez, 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 [15–18].

A fim de encontrar uma medida de complexidade que possa ser aplicada a qualquer série temporal e, nas palavras dos autores, seja também robusta, simples e computacionalmente rápida, Bandt e Pompe [11] 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:

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

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

$$(\vec{2}) = (4,7),$$

$$(\vec{3}) = (7,9),$$

$$(\vec{4}) = (9,10),$$

$$(\vec{5}) = (10,6),$$

$$(\vec{6}) = (6,11),$$

$$(\vec{7}) = (11,3).$$

Em seguida, vamos ordená-los com respeito aos seus valores relativos, obtendo os dois grupos 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("01") = 4/6 e do tipo "10" como p("10") = 2/6.

A entropia de permutação para essa série é, simplesmente, a entropia de Shannon [19] calculada para essas probabilidades. Lembrando que a entropia de Shannon associada a uma distribuição de probabilidades $\{p_1, p_2, \dots, p_m\}$ é¹

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

 $^{^1}$ Ao longo desse texto, vamos considerar o logaritmo para o cálculo da entropia na base e 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 1.18.

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, 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:

$$(\vec{3}) = (4,7,9),$$

$$(\vec{4}) = (7,9,10),$$

$$(\vec{5}) = (9,10,6),$$

$$(\vec{6}) = (10,6,11),$$

$$(\vec{7}) = (6,11,3),$$

os quais podem ser agrupados em 3! = 6 categorias associadas ao ordenamento de seus elementos. Nesse caso, 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 $a_1 < a_0 < a_2$. Aqui, a_i representa o *i*-ésimo elemento dos vetores. Observamos que para d = 3, das 3! permutações possíveis, apenas três estão presentes no exemplo em questão, conduzindo as seguintes probabilidades:

$$p("012") = 0,4,$$

$$p("021") = 0,$$

$$p("102") = 0,2,$$

$$p("120") = 0,$$

$$p("201") = 0,4,$$

$$p("210") = 0,$$

em que $p(\pi_i)$ representa a probabilidade de termos a ocorrência da permutação π_i . Dessa forma, a entropia de permutação para d=3 é dada por

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

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

$$\{x_1, x_2, \dots, x_n\} = \{x_t\}_{t=1,2,\dots,n}.$$
 (1.4)

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

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

com s = d, d + 1, ..., 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, ..., d - 1). Em seguida, determinamos a frequência relativa (probabilidade de ocorrência) de cada uma dessas d! permutações, representada por

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

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

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

Vemos que o único parâmetro usado para determinar a entropia de permutação é d, a chamada embedding dimension. 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 aplicações com séries empíricas, pois para que a estimativa de P seja 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} é obtida 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. Podemos obter a distribuição de probabilidade que maximiza a entropia via método dos multiplicadores de Lagrange [20]. Essa distribuição está sujeita aos vínculos referentes aos axiomas de probabilidade:

$$p(\pi_i) \ge 0 \, (\forall i), \tag{1.9}$$

o qual afirma que a probabilidade de encontrarmos o sistema em qualquer estado π_i é um número real não negativo, e

$$\sum_{i=1}^{d!} p(\pi_i) = 1, \tag{1.10}$$

o qual nos diz que a soma das probabilidades de todos os d! estados acessíveis ao sistema corresponde a 100%. Seguindo o método dos multiplicadores de Lagrange, a função que devemos maximizar sujeita a esses vínculos é dada por

$$\mathcal{L}(p(\pi_i), \lambda) = -\sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i) + \lambda \left\{ \sum_{i=1}^{d!} p(\pi_i) - 1 \right\},$$
 (1.11)

com λ sendo o multiplicador de Lagrange. Em seguida, para encontrarmos os valores extremos de uma função, devemos derivá-la e igualá-la a zero:

$$\frac{\partial \mathcal{L}}{\partial p(\pi_j)} = 0 \implies -\log p(\pi_j) - 1 + \lambda = 0,$$

$$\log p(\pi_j) = \lambda - 1,$$

$$p(\pi_j) = e^{\lambda - 1},$$
(1.12)

ou seja, as probabilidades $p(\pi_j)$ não dependem de j. Realizando a soma para todos os estados, ficamos com

$$\sum_{j=1}^{d!} p(\pi_j) = \sum_{j=1}^{d!} e^{\lambda - 1}.$$
(1.13)

Identificando o do lado esquerdo com o vínculo definido na equação 1.10, escrevemos

$$1 = e^{\lambda - 1} \sum_{j=1}^{d!} 1,$$

$$\frac{1}{d!} = e^{\lambda - 1}.$$
(1.14)

Comparando essa última equação com a última equação em 1.12, concluímos que

$$p(\pi_j) = \frac{1}{d!},\tag{1.15}$$

ou seja, a distribuição de probabilidade "menos informativa" e que maximiza a entropia, é a distribuição uniforme,

$$P_e = \{ p(\pi_i) = 1/d!, i = 1, \dots, d! \}.$$
(1.16)

Calculando a entropia de Shannon para essa distribuição de probabilidade, obtemos

$$S_{max} = S[P_e] = -\sum_{i=1}^{d!} \left(\frac{1}{d!}\right) \log\left(\frac{1}{d!}\right)$$

$$= -\left(\frac{1}{d!}\right) \log\left(\frac{1}{d!}\right) \sum_{i=1}^{d!} 1$$

$$= -\left(\frac{1}{d!}\right) (\log 1 - \log d!) d!$$

$$= \log d!.$$
(1.17)

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

$$H[P] = \frac{S[P]}{S_{max}} = \frac{S[P]}{\log d!}.$$
 (1.18)

Observamos que com a normalização, os valores da entropia ficam limitados ao intervalo $0 \le H[P] \le 1$, 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, em que todas as d! permutações possíveis são equiprováveis.

Para demonstrar a aplicabilidade da entropia de permutação em dados reais, Bandt e Pompe [11] analisam uma série temporal das amplitudes do sinal sonoro obtido a partir da pronúncia da frase "entropia de permutação mede complexidade" (em inglês) para encontrar os segmentos que contenham voz ativa. Essa série é mostrada na figura 1.1(a). A medida que eles utilizam para comparar com a entropia H é a chamada zero-crossing rate (ZCR) [21], que mede a taxa com que a amplitude do sinal muda de positivo para negativo dentro de um intervalo de tempo, ou seja, a taxa local com que a amplitude do sinal passa pelo valor zero. A ZCR é muito usada para detectar segmentos sonoros de curta duração. Ela assume que a energia está concentrada em baixas frequências para a voz ativa [22], o que acarreta em poucas oscilações por unidade de tempo e, portanto, uma contagem baixa de passagens pelo zero. Assim sendo, esperamos um valor baixo para a ZCR quando a voz ativa estiver presente e um valor alto caso contrário.

Na figura 1.1(a), observamos que o sinal começa e termina com ruído, consequentemente, os valores para a entropia H são bem próximos a 1, como mostrado na figura 1.1(b). O mesmo ocorre para os chamados unvoiced sounds, que são sons consonânticos (de baixa intensidade) e também para as pausas durante a fala. Já os segmentos que contêm voz ativa são caracterizados por um decréscimo no valor da entropia. Observamos que as duas técnicas se comportam similarmente, mas é possível notar que a entropia de permutação identifica melhor presença de voz ativa. Por exemplo, durante o silêncio, de 0 a 0,3 segundos ou a partir de 3,6 segundos, percebemos que a ZCR baixa é um indicativo da presença de voz, o que não

é o caso; já a entropia próxima a 1 indica silêncio e/ou ruído. Desse modo, podemos concluir que a entropia de permutação é superior à ZCR, tendo em vista sua melhor capacidade em identificar essas sutilezas, tanto durante a pronúncia, quanto durante as pausas decorrentes das transições.

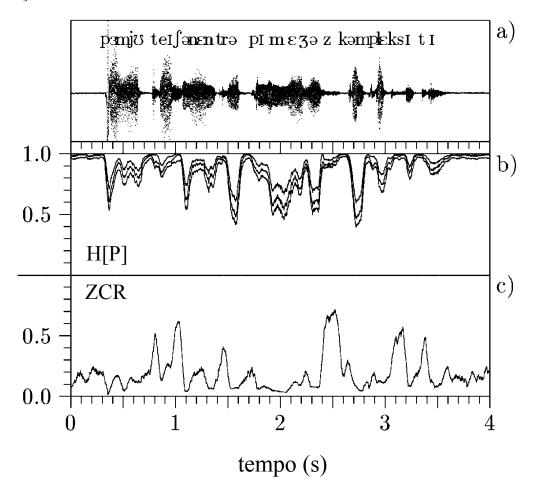


Figura 1.1: (a) Representação da série temporal das amplitudes sonoras ao longo da pronúncia da frase "entropia de permutação mede complexidade" em inglês. Os símbolos acima da série indicam a transcrição fonética da fala. A taxa de amostragem é 11 kHz e a unidade da amplitude é arbitrária. (b) Valor da entropia de permutação normalizada H[P] para d=2,3 e 4 (de cima para baixo). (c) Valor da zero-crossing rate (ZCR). Ambas quantidades foram calculadas dentro de uma janela de tempo móvel com tamanho igual a 46 milissegundos. Figura adaptada da referência [11].

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 um 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 colocações intuitivas apontadas anteriormente. Essa medida pode ser obtida levando em conta o que López-Ruiz et al. [12] definem como "desequilíbrio", \mathcal{D} , que 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, onde os átomos estão completamente 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 minímo, 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 1.2, percebemos que, sozinhas, nem a entropia nem o desequilíbrio são capazes de quantificar essa ideia intuitiva de complexidade estatística para o cristal e para o gás ideal. A solução para esse problema, proposta por López-Ruiz et al. [12], 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.

Desse modo, a partir do trabalho desenvolvido por López-Ruiz et al. [12], obtemos uma medida de complexidade estatística que, agregada a prescrição de Bandt e Pompe, pode trazer à tona informações a respeito da dinâmica dos sistemas que não são capturadas apenas

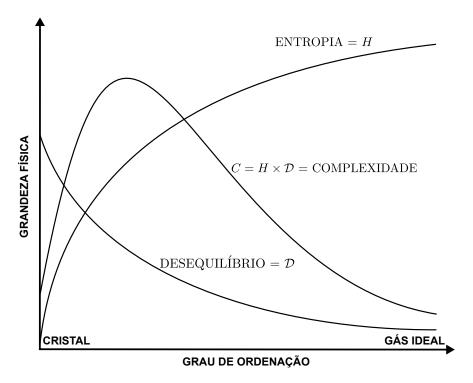


Figura 1.2: Ilustração do comportamento intuitivo esperado para as grandezas entropia H, desequilíbrio \mathcal{D} e complexidade $C = H\mathcal{D}$, todas em função do grau de ordenação do sistema. No extremo de ordem podemos ter um cristal; já no extremo de aleatoriedade, um gás ideal. Notamos que a medida de complexidade estatística de López-Ruiz et~al.~[12] captura a noção intuitiva de complexidade para os exemplos citados, ou seja, tende a zero para ambos. Figura adaptada da referência [12].

pela entropia de permutação. Matematicamente, essa medida é definida como

$$C[P] = H[P]\mathcal{D}[P] = -C_0 \left(\sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i) \right) \left(\sum_{i=1}^{d!} \left(p(\pi_i) - \frac{1}{d!} \right)^2 \right), \tag{1.19}$$

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

Relembrando dos casos extremos ilustrados anteriormente, vemos que essa medida de complexidade é plausível. Para um cristal, o desequilíbrio é \mathcal{D} 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 minímo para um gás ideal, já que os estados são equiprováveis, então $C \to 0$ da mesma forma.

Na verdade, essa interpretação intuitiva de complexidade pode ser obtida também por meio de outras medidas entrópicas e de desequilíbrio existentes na literatura [23, 24]. Além

das já mencionadas entropia de Shannon [19]

$$H[P] = -\sum_{i}^{d!} p(\pi_i) \log p(\pi_i), \tag{1.20}$$

e da distância euclidiana

$$\mathcal{D}[P] = \sum_{i}^{d!} \left(p(\pi_i) - \frac{1}{d!} \right)^2, \tag{1.21}$$

podemos listar:

• Entropia de Tsallis [25]:

$$H_q[P] = \frac{1}{q-1} \left[1 - \sum_{i=1}^{d!} (p(\pi_i))^q \right], \tag{1.22}$$

com q sendo um parâmetro real.

• Entropia de Rényi [26]:

$$H_{\alpha}[P] = \frac{1}{1-\alpha} \log \left[\sum_{i}^{d!} (p(\pi_i))^{\alpha} \right], \quad (\alpha \ge 0 \text{ e } \alpha \ne 1), \tag{1.23}$$

com α sendo um parâmetro real.

• Distância de Wooters [27]:

$$\mathcal{D}[P] = \cos^{-1} \left\{ \sum_{i}^{d!} (p(\pi_i))^{1/2} \left(\frac{1}{d!}\right)^{1/2} \right\}.$$
 (1.24)

• Divergência de Jensen-Shannon [28]:

$$\mathcal{D}[P] = \mathcal{D}_0\{H[(P+P_e)/2] - H[P]/2 - H[P_e]/2\}, \tag{1.25}$$

com H sendo a entropia de Shannon (equação 1.20) e \mathcal{D}_0 uma constante de normalização.

Uma característica interessante da medida de complexidade proposta por López-Ruiz et al. [12] (equação 1.19) é o fato dela não ser uma função unívoca da entropia. 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 1.3. Nessa figura, usamos um exemplo simples para ilustrar esse fato. Por simplicidade, consideramos que o sistema possui somente

3 estados acessíveis. Assim, vamos assumir que a distribuição de probabilidade desses estados seja representada por $P = \{a, b, 1 - (a+b)\}$ com a > 0, b > 0 e $(a+b) \le 1$. 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 1.3.

1.3 Plano complexidade-entropia

O fato descrito na figura 1.3 serviu de motivação para Rosso et al. [13] ao tentar distinguir sinais caóticos e estocásticos. Eles notaram que essa distinção é bastante complicada usando apenas a entropia de permutação. Por conta disso, propuseram o uso de uma representação baseada no trabalho de López-Ruiz et al. [12], na qual duas medidas são avaliadas: a entropia de permutação normalizada H e a complexidade C associada à divergência de Jensen-Shannon (equação 1.25). 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. [13] usaram a abordagem de Bandt e Pompe para extrair a distribuição dos padrões ordinais $P = \{p(\pi_i)\}_{i=1,\dots,d!}$ de várias séries temporais simuladas, cuja dinâmica é conhecida ser caótica ou estocástica. Usando essa distribuição, eles calcularam a entropia de permutação normalizada

$$H[P] = -\frac{1}{\log d!} \sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i), \tag{1.26}$$

e a complexidade estatística

$$C[P] = H[P]\mathcal{D}[P], \tag{1.27}$$

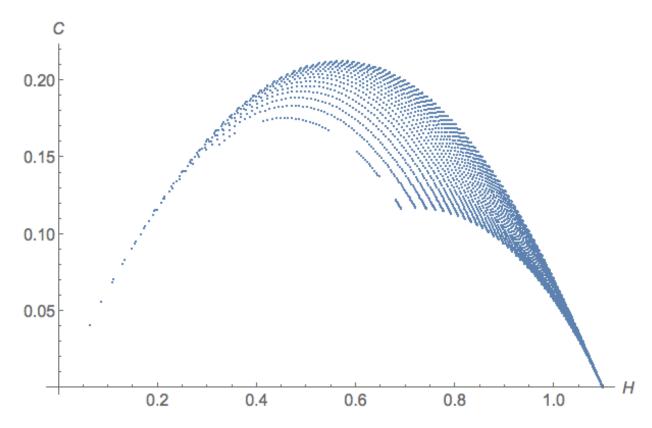


Figura 1.3: Representação do plano complexidade-entropia que ilustra o fato da complexidade não ser uma função unívoca da entropia. Tendo como exemplo o valor de $H \approx 0.6$ obtido para as duas distribuições P_1 e P_2 citadas no texto, observamos que existem vários valores possíveis de C ao combinarmos todos os valores de a e b, tais que a > 0, b > 0 e $(a + b) \le 1$.

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

$$\mathcal{D}[P] = \mathcal{D}_0 \left\{ H \left[\frac{P + P_e}{2} \right] - \frac{H[P]}{2} - \frac{H[P_e]}{2} \right\}. \tag{1.28}$$

Note que

$$\frac{P + P_e}{2} = \left\{ \frac{p(\pi_i) + 1/d!}{2} \right\}_{i=1,\dots,d!}$$
 (1.29)

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, $P^* = \{\delta_{i1}\}_{i=1,\dots,d!}$. Nesse caso, a constante \mathcal{D}_0 fica

$$\mathcal{D}_0 = \left\{ -\frac{1}{2} \left(\frac{d!+1}{d!} \log(d!+1) - \log d! - 2\log 2 \right) \right\}^{-1}.$$
 (1.30)

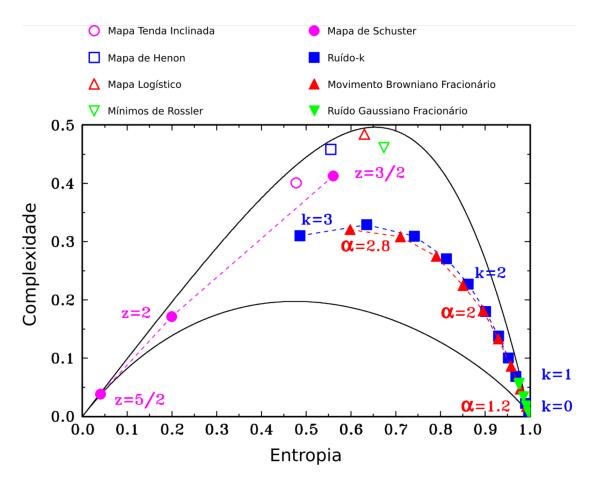


Figura 1.4: As linhas contínuas representam as complexidades mínima C_{min} e máxima C_{max} . A área contida entre elas delimita o plano complexidade-entropia. Os símbolos nesse gráfico indicam a localização de diferentes sinais caóticos (os cinco primeiros da legenda) e estocásticos (os três últimos) nesse plano. O mapa de Schuster é caracterizado por sinais caóticos intermitentes que possuem ruído na forma $1/f^z$, os ruídos-k têm a mesma representação, ou seja, $1/f^k$. O movimento browniano fracionário e o ruído browniano fracionário também são caracterizados por um espectro de potência da forma $1/|f|^{\alpha}$, com a relação: $\alpha = 2h + 1$ ($1 < \alpha < 3$) para o movimento browniano fracionário e $\alpha = 2h - 1$ ($-1 < \alpha < 1$) para o ruído browniano fracionário, com h sendo o expoente de Hurst, 0 < h < 1. Percebemos que esses sinais são facilmente distinguíveis, mesmo quando possuem a mesma entropia. Por exemplo, séries estocásticas do ruído-k podem ter a mesma entropia de séries caóticas do mapa logístico, porém complexidades completamente diferentes. Figura adaptada da referência [13].

A figura 1.4, mostra o plano complexidade-entropia obtido por Rosso et~al. para diferentes séries temporais de natureza estocástica e caótica. Nessa figura, com exceção dos círculos preenchidos, os demais símbolos preenchidos mostram os valores de H e C para séries estocásticas. Já os símbolos não preenchidos representam séries temporais caóticas. Podemos observar que as séries estocásticas e caóticas ocupam regiões distintas do plano complexidade-entropia, de modo a serem facilmente distinguíveis. Vale notar ainda que, sozinhas, a entropia de permutação ou a complexidade estatística não são capazes de fazer

essa distinção. Por exemplo, séries temporais do movimento browniano fracionário podem apresentar a mesma entropia de séries do mapa logistíco, mas complexidades diferentes. Similarmente, o mapa de Schuster para z=2 tem complexidade similar a do movimento browniano fracionário, mas o valor da entropia é muito diferente. De maneira geral, Rosso et al. [13] observaram que sistemas caóticos estão localizados próximos à região de máxima complexidade. Segundo argumentam, isso ocorre porque esses sistemas possuem um grau maior de estrutura em suas dinâmicas, como a ausência de certos padrões ordinais. Assim, vemos que a complexidade estatística quantifica, além da aleatoriedade, esse grau de estruturas existentes.

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. [14] propuseram uma extensão para sistemas bidimensionais (ou de dimensão maior) do plano complexidade-entropia. Nesse trabalho, eles verificam a utilidade do método por meio dos seguintes exemplos:

- distinção de imagens de superfícies fractais com dimensões fractais distintas;
- identificação de transições de fase em cristais líquidos a partir de imagens de suas texturas;
- identificação de transições de fase no modelo de Ising, a partir de imagens das chamadas superfícies de Ising.

Seguindo a ideia do início desse capítulo, vamos exemplificar a extensão do método proposto por Ribeiro *et al.* antes de apresentá-lo 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 simboliza um *pixel* dessa imagem. Vamos considerar, por simplicidade, a seguinte matriz de tamanho 3×3 :

$$A = \left(\begin{array}{ccc} 3 & 4 & 8 \\ 5 & 6 & 7 \\ 2 & 8 & 9 \end{array}\right).$$

Analogamente ao vetor \vec{s} (definido na equação 1.5), que representa partições na série temporal, no caso bidimensional temos matrizes superpostas de tamanho $d_x \times d_y$ ($d_x, d_y > 1$).

Por exemplo, para $d_x = d_y = 2$, obtemos quatro submatrizes da forma

$$A_i = \left(\begin{array}{cc} a_0 & a_1 \\ a_2 & a_3 \end{array}\right)$$

diferentes associadas à A, que 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: $a_0 \le a_1 \le a_2 \le a_3$. A submatriz A_2 , por sua vez, corresponde à sequência "0231", já que $a_0 \le a_2 \le a_3 \le a_1$. Finalmente, A_3 é associada ao padrão "2013" pois $a_2 \le a_0 \le a_1 \le a_3$. De posse das permutações, calculamos as probabilidades de ocorrência $p(\pi_i)$ de cada uma:

$$p("0123") = 0.50,$$

 $p("0231") = 0.25,$
 $p("2013") = 0.25.$

Vale notar que de todas as $(d_x d_y)! = 24$ permutações possíveis, apenas as três anteriores surgiram nesse exemplo, as demais são tratadas com probabilidade de ocorrência nula. Usando o conjunto de probabilidades $P = \{p(\pi_i)\}_{i=1,\dots,(d_x d_y)!}$, 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. [14], vamos definí-la formalmente. Para isso consideremos uma matriz bidimensional $\{y_i^j\}_{i=1,\dots,n_x}^{j=1,\dots,n_y}$ de tamanho $n_x \times n_y$. As submatrizes (s_x, s_y) de tamanho $d_x \times d_y$ $(d_x, d_y > 1)$, que aqui fazem o papel do vetor

 (\vec{s}) , são dadas por

$$(s_{x}, s_{y}) \mapsto \begin{pmatrix} y_{s_{x}-(d_{y}-1)}^{s_{y}-(d_{y}-1)} & y_{s_{x}-(d_{x}-2)}^{s_{y}-(d_{y}-1)} & \dots & y_{s_{x}-1}^{s_{y}-(d_{y}-1)} & y_{s_{x}}^{s_{y}-(d_{y}-1)} \\ y_{s_{x}-(d_{x}-1)}^{s_{y}-(d_{y}-2)} & y_{s_{x}-(d_{x}-2)}^{s_{y}-(d_{y}-2)} & \dots & y_{s_{x}-1}^{s_{y}-(d_{y}-2)} & y_{s_{x}}^{s_{y}-(d_{y}-2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{s_{x}-(d_{x}-1)}^{s_{y}-1} & y_{s_{x}-(d_{x}-2)}^{s_{y}-1} & \dots & y_{s_{x}-1}^{s_{y}-1} & y_{s_{x}}^{s_{y}-1} \\ y_{s_{x}-(d_{x}-1)}^{s_{y}-1} & y_{s_{x}-(d_{x}-2)}^{s_{y}-1} & \dots & y_{s_{x}-1}^{s_{y}-1} & y_{s_{x}}^{s_{y}} \end{pmatrix},$$

$$(1.31)$$

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

$$\pi_i = [(r_0, u_0), (r_1, u_0), \dots, (r_{d_x - 1}, u_0), \dots, (r_0, u_{d_y - 1}), (r_1, u_{d_y - 1}), \dots, (r_{d_x - 1}), u_{d_y - 1})], \quad (1.32)$$

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

$$y_{s_x-r_{d_x-1}}^{s_y-u_{d_y-1}} \le y_{s_x-r_{d-2}}^{s_y-u_{d_y-1}} \le \dots \le y_{s_x-r_1}^{s_y-u_{d_y-1}} \le y_{s_x-r_0}^{s_y-u_{d_y-1}} \le \dots$$

$$\le y_{s_x-r_{d_x-1}}^{s_y-u_0} \le y_{s_x-r_{d-2}}^{s_y-u_0} \le \dots \le y_{s_x-r_1}^{s_y-u_0} \le y_{s_x-r_0}^{s_y-u_0}.$$

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

$$p(\pi_i) = \frac{\#\{(s_x, s_y) | s_x \le n_x - d_x + 1 \text{ e } s_y \le n_y - d_y + 1; \ (s_x, s_y) \text{ \'e do tipo } \pi_i\}}{(n_x - d_x + 1)(n_y - d_y + 1)}.$$
 (1.33)

Analogamente ao caso unidimensional, os únicos parâmetros que precisamos definir 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

$$H[P] = \frac{S[P]}{S_{max}} = \frac{S[P]}{\log[(d_x d_y)!]},$$
(1.34)

na qual S[P] é a entropia de Shannon da distribuição de probabilidades P (equação 1.20) 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

$$C[P] = \mathcal{D}[P]H[P], \tag{1.35}$$

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

$$\mathcal{D}[P] = \mathcal{D}_0 \left\{ H \left[\frac{P + P_e}{2} \right] - \frac{H[P]}{2} - \frac{H[P_e]}{2} \right\}, \tag{1.36}$$

com

$$\mathcal{D}_0 = \left\{ -\frac{1}{2} \left(\frac{(d_x d_y)! + 1}{(d_x d_y)!} \log[(d_x d_y)! + 1] - 2 \log[2(d_x d_y)!] + \log[(d_x d_y)!] \right) \right\}^{-1}$$
 (1.37)

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.

Como mencionado anteriormente, Ribeiro et al. fazem uma série de aplicações em dados bidimensionais para verificar a utilidade do método. Iremos apresentar apenas uma dessas aplicações relacionadas às superfícies de Ising [29]. Essas superfícies são obtidas pela acumulação dos valores dos spins do modelo de Ising definido pela hamiltoniana

$$\mathcal{H} = -\sum_{\langle i,j \rangle} \sigma_i \sigma_j,\tag{1.38}$$

na qual a soma é sobre todos os pares de primeiros vizinhos na rede. Resolvendo numericamente o modelo de Ising para spin 1/2 ($\sigma_i \in \{-1,1\}$) em uma rede de tamanho $L \times L$ com condições de contorno periódicas via método de Monte Carlo [30], Ribeiro et al. definem a altura da superfície associada a cada sítio i da rede como

$$S_i = \sum_t \sigma_i(t), \tag{1.39}$$

com t representando o número de passos do algoritmo de Monte Carlo.

Na figura 1.5, observamos três superfícies obtidas por esse método para diferentes valores de temperatura reduzida T/T_c , sendo $T_c = 2/\ln(1+\sqrt{2})$ a temperatura crítica do modelo de Ising. Abaixo e acima da temperatura crítica T_c notamos um padrão aleatório nas imagens, já na temperatura crítica de transição de fase percebemos um padrão mais complexo.

Na figura 1.6, observamos a dependência das medidas de complexidade H e C com a temperatura reduzida T/T_c , após um grande número de passos de Monte Carlo para uma rede quadrada de tamanho L=500. As figuras 1.6(a) e 1.6(b) mostram H e C para $d_x=2$ e $d_y=3$ e para a rotação $d_x\to d_y$ e $d_y\to d_x$. Claramente, notamos um pico/vale bem acentuado na temperatura crítica T_c em ambas as medidas. Além disso, observamos que H e C são praticamente invariantes perante a rotação das submatrizes (s_x,s_y) . Já na figura 1.6(c), temos uma visualização tridimensional da transição de fase para $d_x=d_y=3$. Esse tipo de representação é útil pois pode tornar o ponto crítico mais visível em transições de fase mais complexas.

Com esse exemplo, encerramos a apresentação da técnica de entropia e complexidade de permutação. Nos próximos capítulos, vamos nos dedicar ao tema central dessa dissertação: a caracterização de obras de arte por meio dessas medidas de complexidade estatística.

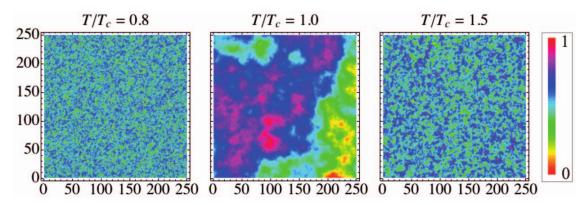


Figura 1.5: Exemplos de superfícies de Ising. Essas superfícies foram obtidas após 10^5 passos de Monte Carlo para três temperaturas diferentes: abaixo de T_c , em T_c e acima de T_c , com T_c representando a temperatura crítica de transição de fase do modelo de Ising. Os valores da altura estão escalados de forma a permanecerem entre 0 e 1. Notamos que para temperaturas maiores ou menores que T_c , essas superfícies exibem um padrão praticamente aleatório, devendo possuir um valor baixo para a complexidade. Para valores próximos a T_c , a superfície mostra um padrão mais complexo, indicando também um aumento no valor da medida de complexidade. Figura extraída da referência [14].

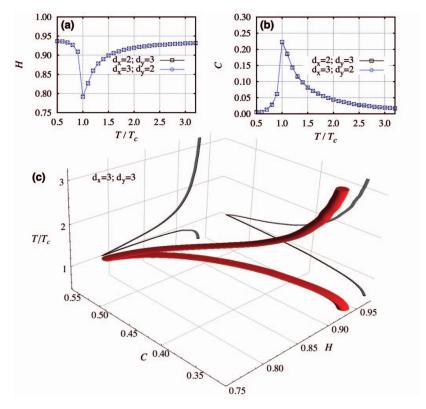


Figura 1.6: Dependência das medidas de complexidade com a temperatura reduzida T/T_c para as superfícies de Ising. (a) A entropia de permutação H e (b) a complexidade estatística C versus a temperatura reduzida T/T_c para $d_x=2$ e $d_y=3$ e também para a rotação $d_x=3$ e $d_y=2$. Notamos claramente o ponto crítico da transição de fase. (c) Visualização tridimensional da transição de fase para $d_x=d_y=3$. Em cinza temos a dependência da temperatura reduzida T/T_c com H e C. Figura extraída da referência [14].

A base de dados sobre artes visuais Wiki Art

Neste capítulo, apresentamos como os dados utilizados para as análises foram obtidos e processados. Esses dados consistem de imagens digitais de um grande conjunto (quase 140 mil imagens) de obras de arte e informações relacionadas às obras, como a data, gênero e estilo. Após manipuladas, cada uma dessas imagens é convertida numa matriz, a qual é utilizada para calcular os valores de entropia e complexidade de permutação.

2.1 Descrição e apresentação dos dados

As imagens digitais das obras de arte utilizadas para nossas análises foram obtidas a partir de uma das maiores bases de dados *online* sobre artes visuais disponível gratuitamente: a *Wiki Art* (www.wikiart.org). Além das imagens, essa base contém vários metadados relacionados às pinturas, como data em que foram produzidas, estilo a que pertencem (Impressionismo, Surrealismo, Cubismo, etc.) e o gênero (retrato, paisagem, abstrato, etc.).

Em agosto de 2016, quando fizemos o download das imagens, essa base continha 137 364 arquivos que totalizam 27,3 GB de informação. De todos esse arquivos, removemos de nossas análises as obras que não apresentavam data e também imagens associadas à esculturas, fotografias, caricaturas, grafites e outras não relacionadas à pintura artistíca. Essa filtragem foi realizada a partir de metadados disponibilizados na base Wiki Art. Dessa forma, nossas análises foram realizadas em 96 685 imagens de 2 424 artistas diferentes. Essa quantidade é bastante significativa, tendo em vista que a maioria dos trabalhos existentes na literatura sobre aspectos das artes visuais se restringem apenas a artistas ou pinturas específicas [31–35].

Inicialmente, faremos uma descrição demográfica de nossos dados. Na figura 2.1, mostra-

mos como as pinturas estão distribuídas ao longo do tempo, ou seja, o número de pinturas por ano. Notamos que a maior parte das obras foram produzidas após a primeira metade do século XIX. Após uma inspeção em nossa base de dados, verificamos que existem três estilos predominantes presentes no período anterior a 1800, a Renascença e suas vertentes (como a inicial, alta, tardia e no norte), o Barroco e o Rococó. Esses três estilos juntos somam aproximadamente 10 mil imagens.

Selecionamos os estilos que possuem mais de mil pinturas e, nas figuras 2.2 e 2.3, mostramos a fração das pinturas por estilo ao longo do tempo, deixando claro o fato mencionado anteriormente e ilustrando o comportamento para os demais estilos. Na figura 2.3, mostramos a fração das pinturas por estilo a partir de 1800, já que a maioria dos movimentos artísticos foram desenvolvidos após esse período.

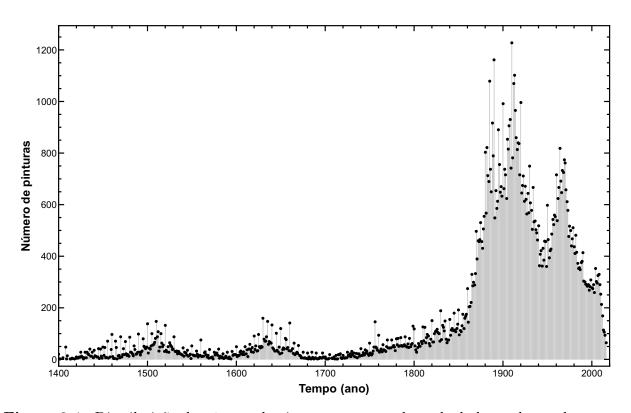


Figura 2.1: Distribuição do número de pinturas em nossa base de dados ao longo do tempo. Aqui, cada ponto representa o número de imagens disponíveis para um dado ano.

Na figura 2.4, mostramos o número total de pinturas para cada um dos 157 estilos presentes em nossa base de dados. É interessante mencionar que os chamados movimentos vanguardistas¹ juntos somam aproximadamente 50 mil imagens.

¹Os movimentos vanguardistas são aqueles que se desenvolveram nas décadas iniciais do século XX, que apresentavam como principal caracteristíca a oposição ao tradicionalismo do século XIX, causando uma ruptura cultural de modelos preestabelecidos. Os principais incluem Cubismo, Expressionismo, Dadaísmo e Surrealismo.

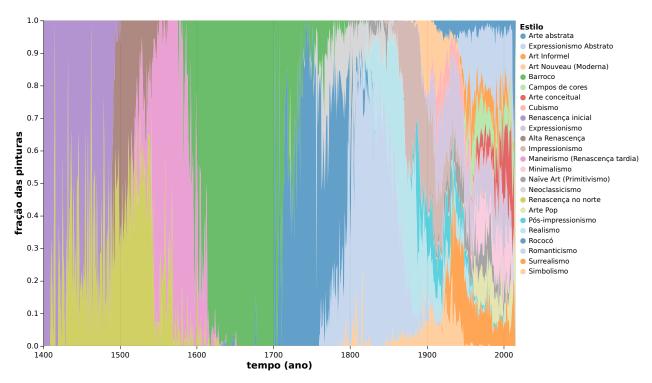


Figura 2.2: Fração de ocorrência das pinturas por estilo ao longo do tempo. Apenas os estilos que possuem mais de mil pinturas estão presentes. Nesse gráfico, cada cor representa um estilo e, para um dado ano, a "extensão vertical da cor" representa a fração de ocorrência do estilo correspondente.

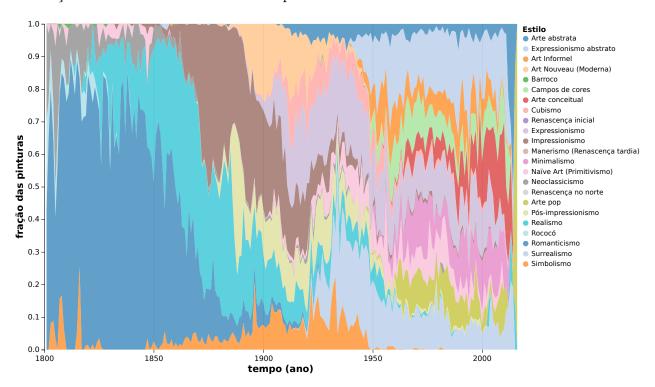


Figura 2.3: O mesmo gráfico da figura 2.2 ao considerar apenas as pinturas produzidas após 1800.

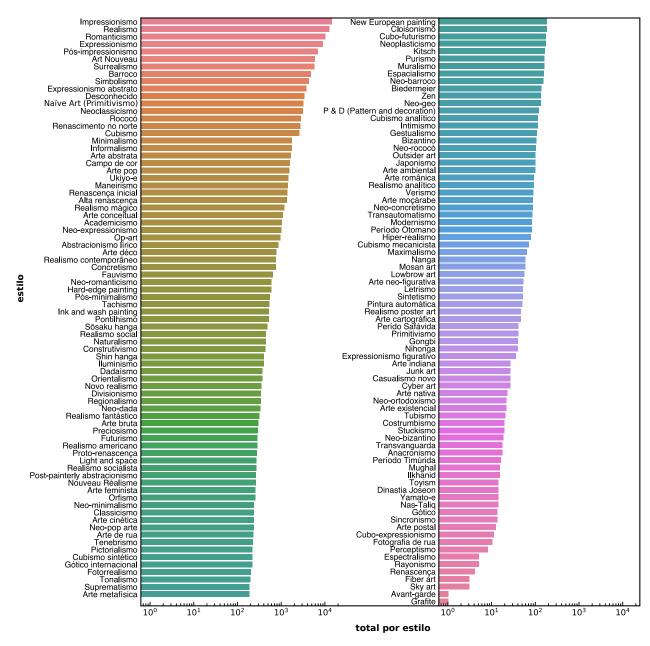


Figura 2.4: Distribuição das pinturas por estilo artístico. Aqui, mostramos todos os 157 estilos presentes em nossa base de dados e o número total de pinturas para cada um.

2.2 Utilizando a linguagem de programação *Python* e seus pacotes para extração e processamento dos dados

Como já foi mencionado, em agosto de 2016, a base Wiki Art continha 137 364 imagens, um número expressivo que praticamente impossibilita realizar o download manualmente. Por conta disso e também para estudar ferramentas computacionais disponíveis para abordar os

sistemas complexos, os quais usualmente são associados a uma grande quantidade de dados, utilizamos pacotes em linguagem de programação *Python* [36], como o *Scrapy* [37], que tem por objetivo extrair dados de páginas da *internet* de forma automatizada, otimizando substancialmente o tempo utilizado nesse processo.

Na figura 2.5, mostramos um dos códigos utilizados para coletar dados da base *Wiki Art*. Em particular, esse código acessa a página mostrada na figura 2.6 que contém todos os artistas por letra inicial do nome. Ao percorrer todo o alfabeto, obtemos como resultado uma lista contendo todos os artistas. Cada linha dessa lista, contém o *link* para a página do artista e seu nome, como também mostramos na figura 2.6.

```
import scrapy
   import string
   from scrapy.http import Request
   from scrapy.selector import HtmlXPathSelector
   class WikiartItem(scrapy.Item):
     # aqui definimos os campos que iremos extrair, nome do artista e
    link que direciona para a página que contém suas obras.
       name = scrapy.Field()
       link = scrapy.Field()
   class Wikiart(scrapy.Spider):
     # a spider é o objeto que irá percorrer a paginá e extrair os dados
     , definimos em quais páginas ela deve rodar, iremos alterando apenas
     o último caractere da paginá, passando por todo o alfabeto.
       name = 'wikiart'
       allowed_domains = ['wikiart.org']
14
       start_urls = ['http://www.wikiart.org/en/Alphabet/%s' %s for s in
     list(map(chr, range(ord('A'), ord('Z')+1)))]
        start_urls.append("http://www.wikiart.org/en/Alphabet/%C3%B8")
      # a função abaixo é definida na spider, é ela quem, finalmente,
17
     extrai os dados.
       def parse(self, response):
            hxs = HtmlXPathSelector(response)
19
            sites = hxs.select('//li[@class="title"]')
            artists = []
            for site in sites:
                artist = WikiartItem()
23
                artist['name'] = site.select('a/text()')[0].extract()
                artist['link'] = site.select('a/@href')[0].extract()
                artists.append(artist)
            return artists
```

Figura 2.5: Código em linguagem de programação *Python* [36] utilizando o pacote *Scrapy* [37] para realizar a extração automatizada de dados da página mostrada na figura 2.6.

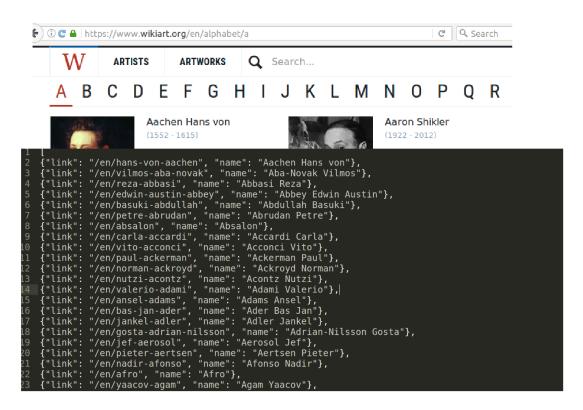


Figura 2.6: Ilustração da página *Wiki Art* contendo todos os artistas por letra inicial do nome e em ordem alfabética. O código mostrado na figura 2.5 lê essas páginas para extrair os *links* para a página individual de cada artista e seus nomes, conforme também está ilustrado nessa figura.

Após obter os artistas e os *links* para suas páginas, acessamos, também de maneira automatizada, cada uma delas. Nas páginas dos artistas, encontramos todas as pinturas disponíveis. Por sua vez, cada pintura tem uma página individual associada, na qual encontramos vários metadados, como data, estilo, genêro, título, mídia, dimensão e o *link* para download, como mostrado na figura 2.7. Ao percorrer todos os artistas, construímos um arquivo de texto com 137 364 linhas (uma para cada imagem), conforme também está ilustrado na figura 2.7. De posse desse arquivo, em particular de todos os *links* das imagens, utilizamos um outro programa em *Python* para executar o *download*.

O próximo passo é processar cada imagem para calcularmos as medidas de complexidade H e C. Na figura 2.8, ilustramos como esse processo é realizado. Para o computador a imagem é interpretada como uma matriz de pixels. Em nosso exemplo, a obra Abaporu de Tarsila do Amaral tem tamanho 40×47 pixels. No caso do padrão RGB, cada pixel é subdivido em três camadas, uma vermelha (R), uma verde (G) e uma azul (B). A intensidade de cada componente R, G e B de um pixel varia de 0 a 255, sendo esse o intervalo de valores dos elementos das matrizes que representam as camadas. A matriz com a soma das intensidades das três camadas, pixel a pixel, é a que utilizamos para calcular as medidas de complexidade H e C.

Na figura 2.9, observamos como esse processo pode ser implementado utilizando o pacote

OpenCV [38], também em linguagem Python [36]. Nesse código, observamos uma parte responsável por executar outro código, escrito em linguagem de programação C (por questão de eficiência computacional), que lê a matriz da soma das três camadas e calcula a entropia e a complexidade de permutação para diferentes valores de d_x e d_y . Esse código também calcula as dimensões da matriz que representa a imagem, ou seja, o tamanho da imagem.

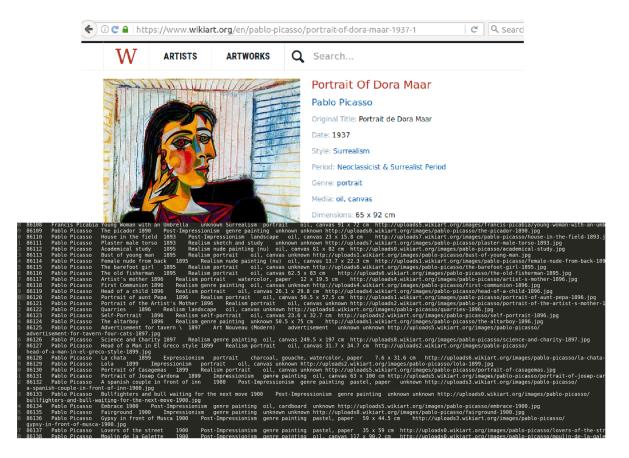


Figura 2.7: Em segundo plano, mostramos uma ilustração de uma página do *Wiki Art* contendo os metadados de uma pintura. Em primeiro plano, as informações obtidas para algumas imagens.

De posse de todas essas informações, construímos um arquivo contendo metadados relacionados às pinturas, as medidas de entropia H e complexidade de permutação C e também o tamanho da imagem. Uma ilustração do arquivo final que foi utilizado para todas as análises que serão apresentadas no próximo capítulo é mostrada na figura 2.10.

Por fim, a figura 2.11 ilustra o tipo de informação que as medidas de complexidade H e C podem fornecer sobre as pinturas. Nessa figura, mostramos três imagens, que fazem uma analogia aos sistemas físicos da figura 1.2. A imagem a esquerda apresenta um padrão bem regular e é caracterizada por $H\approx 0$ e $C\approx 0$, como o cristal. Já a imagem a direita, apresenta um padrão bem mais irregular ou aleatório, sendo caracterizada por $H\approx 1$ e $C\approx 0$, analogamente ao gás ideal. A imagem central representa um caso intermediário, no qual $H\neq 0$ e $H\neq 1$, assim como $C\neq 0$.

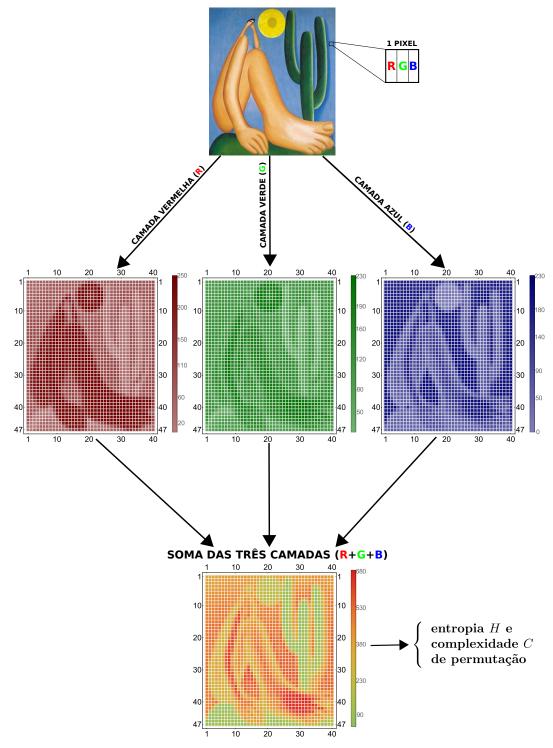


Figura 2.8: Ilustração do procedimento utilizado para processar as imagens e calcular a entropia e complexidade de permutação. Para o computador, a imagem é interpretada como uma matriz de *pixels*, que são subdivididos em três camadas RGB, vermelha (R), verde (G) e azul (B). A imagem acima, *Abaporu* de Tarsila do Amaral, possui tamanho 40×47 *pixels* (usamos uma imagem reduzida apenas para fins ilustrativos). A intensidade de cada camada R, G e B de cada pixel varia de 0 a 255, conforme mostram as figuras intermediárias. A figura mais abaixo mostra a matriz da soma das três camadas, a partir da qual calculamos os valores da entropia H e da complexidade C de permutação.

```
import cv2 # importando o pacote OpenCV
    import numpy as np
    import os
    for i in range(1,137365):
5
          print i
          try:
        # nome do arquivo da imagem, vai de 1.jpg até 137364.jpg.
                  img_file = str(i) +'.jpg'
        # leitura da imagem pelo pacote open cv, o resultado são três
     matrizes correspondentes às camadas rgb (red, green, blue).
                  img = cv2.imread(img_file, cv2.IMREAD_COLOR)
        # soma das matrizes que representam as camadas, resultando em uma
     matriz.
                  imgrgbtotal= (1.)*img[:,:,0]+(1.)*img[:,:,1]+(1.)*img
13
     [:,:,2]
        # salvando a matriz com o total em um arquivo temporário ".temp"
                  np.savetxt(str(i)+'.temp', imgrgbtotal, fmt='%d',
15
     delimiter='\t')
16
        # executa o código em linguagem C, bppprob2dn, que processa a
17
    matriz ".temp" e calcula a entropia e complexidade de permutação. Os
      argumentos 2 2 referem-se aos valores de dx e dy.
18
        os.system("bpprob2dn " + str(i) + ".temp " + str(img.shape[1]) +
19
     " " + str(img.shape[0]) + " 2 2 >> " + str(i) + ".bp")
        # remove o arquivo temporário
                  os.system("rm " + str(i) +".temp")
        # salva o tamanho da imagem
23
        print(img.shape[0],img.shape[1])
24
          except:
25
                  print "error when evaluating " + str(i)
26
```

Figura 2.9: Código em linguagem *Python* [36] utilizando o pacote *OpenCV* [38] para realizar o processamento das imagens, conforme ilustrado na figura 2.8.

• index	\$	artist 🕈	title ≑	date 🕈	style 🗢	genre 🕈	media 🗢					
0	1	Hans von Aachen	Self-portrait	1574.0	Mannerism (Late Renaissance)	unknown	unknown					
1	2	Hans von Aachen	Two laughing men (double self-portrait)	1574.0	Mannerism (Late Renaissance)	unknown	unknown					
2	3	Hans von	Augustus and the	1580.0	*	dimensions		url \$	H22 \$	C22 \$	sizex 🕈	sizey 🗢
			•		· · · · · · · · · · · · · · · · · · ·	unknow	n upload		0.891762	0.131570	423	600
3	4			1585.0	,			aachen/se.				
4	5	Hans von	Portrait of Joseph Heintz	1585.0	Mannerism (Late		n upload	s1.wikiart.org/images/hans-von- aachen/tw	0.893904	0.133453	472	600
		Aachen			Renaissance)	unknow	n upload	s5.wikiart.org/images/hans-von- aachen/au	0.970732	0.038636	542	640
						unknow	n upload	s1.wikiart.org/images/hans-von- aachen/po	0.699617	0.240248	391	500
						unknow	n upload	s7.wikiart.org/images/hans-von- aachen/po	0.897553	0.126936	387	500
	0 1 2 3	0 1 1 2 2 3 3 4	0 1 Hans von Aachen 1 2 Hans von Aachen 2 3 Hans von Aachen 3 4 Hans von Aachen	0 1 Hans von Aachen Self-portrait 1 2 Hans von Aachen (double self-portrait) 2 3 Hans von Aachen Tiburtine Sibyl 3 4 Hans von Aachen Billiverti 4 5 Hans von Portrait of Joseph Heintz	0 1 Hans von Aachen Self-portrait 1574.0 1 2 Hans von Aachen Two laughing men (double self-portrait) 1574.0 2 3 Hans von Aachen Augustus and the Tiburtine Sibyl 1580.0 3 4 Hans von Aachen Portrait of Jacopo Biliverti 1585.0 4 5 Hans von Portrait of Joseph Heintz 1585.0	0 1 Hans von Aachen Self-portrait 1574.0 Mannerism (Late Renaissance) 1 2 Hans von Aachen Two laughing men (double self-portrait) 1574.0 Mannerism (Late Renaissance) 2 3 Hans von Aachen Augustus and the Tiburtine Sibyl 1580.0 Mannerism (Late Renaissance) 3 4 Hans von Aachen Portrait of Jacopo Biliverti 1585.0 Mannerism (Late Renaissance) 4 5 Hans von Portrait of Joseph Heintz 1585.0 Mannerism (Late Renaissance)	1	1	1 Hans von Aachen Self-portrait 1574.0 Mannerism (Late Renaissance) 1 2 Hans von Aachen (double self-portrait) 2 3 Hans von Aachen Tiburtine Sibyl Renaissance) 3 4 Hans von Aachen Biliverti 1585.0 Mannerism (Late Renaissance) 4 5 Hans von Aachen Renaissance) 4 5 Hans von Aachen Aachen Later Renaissance) 6 Wannerism (Later Renaissance) 7 Wannerism (Later Renaissance) 8 Wannerism (Later Renaissance) 9 Wannerism (Later Renaissance) 1585.0 Mannerism (Later Renaissance) 1585.0 Wannerism (Lat	1 Hans von Aachen Self-portrait 1574.0 Mannerism (Late Renaissance) 1 2 Hans von Aachen (double self-portrait) 2 3 Hans von Aachen Tiburtine Sibyl Renaissance) 3 4 Hans von Aachen Billiverti Renaissance) 4 5 Hans von Portrait of Joseph Heintz Aachen Aachen Hans von Aachen United Field Fi	0 1 Hans von Aachen Self-portrait 1574.0 Mannerism (Late Renaissance) unknown unknown 1 2 Hans von Aachen Two laughing men (double self-portrait) 1580.0 Mannerism (Late Renaissance) unknown unknown 2 3 Hans von Aachen Augustus and the Tiburtine Sibyl 1580.0 Mannerism (Late Renaissance) dimensions ≠ url ≠ H22 ≠ C22 ≠ 3 4 Hans von Aachen Portrait of Jacopo Biliverti 1585.0 Mannerism (Late Renaissance) unknown uploads1.wikiart.org/mages/hans-von-aachen/w 0.893904 0.133453 4 5 Hans von Aachen Portrait of Joseph Heintz Park Indian Park In	0 1 Hans von Aachen Self-portrait 1574.0 Mannerism (Late Renaissance) unknown unknown 1 2 Hans von Aachen Two laughing men (double self-portrait) 1574.0 Mannerism (Late Renaissance) unknown unknown 2 3 Hans von Aachen Augustus and the Tiburtine Sibyl 1580.0 Mannerism (Late Renaissance) unknown uploads4.wikiart.org/images/hans-von-aachen/se 0.891762 0.131570 423 3 4 Hans von Aachen Portrait of Jacopo Biliverti 1585.0 Mannerism (Late Renaissance) unknown uploads1.wikiart.org/images/hans-von-aachen/w 0.893040 0.133453 472 4 5 Hans von Aachen Portrait of Joseph Heintz 1585.0 Mannerism (Late Renaissance) unknown uploads1.wikiart.org/images/hans-von-aachen/w 0.990703 0.038636 542 4 5 Hans von Aachen Portrait of Joseph Heintz 1585.0 Mannerism (Late Renaissance) unknown uploads1.wikiart.org/images/hans-von-aachen/w 0.9970732 0.038636 542 4 5 Hans von Aachen Portrait of Joseph Heintz 1585.0 <th< th=""></th<>

Figura 2.10: Ilustração do arquivo final que contém os dados utilizados para as análises que serão apresentadas no próximo capítulo.



Figura 2.11: Ilustração do resultado esperado ao aplicar as medidas de entropia (H) e complexidade (C) de permutação às imagens das pinturas. Fazendo uma analogia com o cristal e o gás ideal (figura 1.2), observamos que a pintura a esquerda faz o papel de um cristal, ou seja, é bastante ordenada e caracterizada por $H \approx 0$ e $C \approx 0$. Já a pintura a direita, é análoga ao gás ideal, ou seja, é bastante desordenada e caracterizada por $H \approx 1$ e $C \approx 0$. A pintura ao centro representa um sistema intermediário, para o qual $H \neq 0$ e $C \neq 0$. Vale mencionar que, embora as imagens a esquerda e a direita sejam caracterizadas por valores extremos de entropia, ambos apresentam complexidade nula, assim como o cristal e o gás ideal.

Plano complexidade-entropia aplicado às imagens de obras de arte

Neste capítulo, apresentamos os resultados das análises que foram realizadas nos dados obtidos da base Wiki Art. Verificamos que utilizando o plano complexidade-entropia conseguimos diferenciar os diversos estilos artísticos, bem como agrupá-los de forma hierárquica de acordo com suas distâncias relativas nesse plano. Além disso, observamos uma evolução temporal clara no desenvolvimento dos movimentos artistícos ao longo da história da arte.

3.1 Independência das medidas de complexidade estatística com a dimensão das imagens

Começamos investigando o efeito do tamanho da imagem nas medidas de complexidade estatística H e C. Essa questão é importante para garantir que os valores de H e C não apresentam algum viés com o tamanho da imagem. Desse modo, podemos nos assegurar que a localização no plano complexidade-entropia reflete as características artísticas da pintura e não apenas suas dimensões. Na figura 3.1(a), temos um histograma que mostra como as imagens estão distribuídas de acordo com uma dimensão típica da imagem: a raiz quadrada de sua área. Dessa figura, notamos que o tamanho das imagens está distribuído ao redor de seu valor médio, que é igual a 890 pixels, com desvio padrão igual a 547 pixels. Observamos, ainda, que essa distribuição é relativamente bem localizada/concentrada, em particular, 93 604 imagens possuem dimensão típica entre $400 \ pixels$ e $1000 \ pixels$.

Para verificarmos se a dimensão da imagem exerce algum efeito nas medidas de complexidade estatística, construímos as figuras 3.1(b) e 3.1(c), em que no eixo horizontal está a dimensão típica da imagem e no eixo vertical as medidas de complexidade estatís-

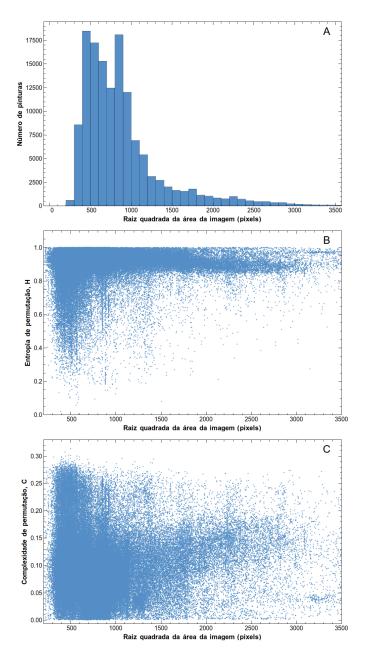


Figura 3.1: As medidas de complexidade não são linearmente correlacionadas com o tamanho das imagens. (A) Histograma mostrando como as imagens estão distribuídas de acordo com a raiz quadrada de sua área, que representa uma dimensão típica da imagem. Notamos que o tamanho das imagens está distribuído ao redor de seu valor médio (890 pixels, com desvio padrão igual a 547 pixels). Exatamente, 93 604 imagens (68,1 % da base de dados) possuem dimensão típica entre 400 pixels e 1000 pixels. (B) Gráfico de dispersão da dimensão típica da imagem pela entropia de permutação H, com cada ponto representando uma imagem. (C) Mesmo gráfico que o anterior, agora para a complexidade de permutação C. Ao calcularmos o coeficiente de Pearson (ρ) , que quantifica o grau de correlação linear entre as variáveis, obtemos $\rho \approx 0.05$ para a relação entre o tamanho da imagem e a entropia de permutação H e $\rho \approx 0.01$ para a relação com a complexidade de permutação C, confirmando a ausência de correlação linear significativa entre as medidas de complexidade e a dimensão da imagem.

tica (entropia e complexidade de permutação, respectivamente). Cada ponto nesses gráficos de dispersão representa uma imagem. Visualmente não percebemos nenhuma relação clara entre as medidas de complexidade e o tamanho das imagens. Para confirmar esse indício, calculamos o coeficiente correlação de Pearson (ρ) [39]. Esse coeficiente quantifica o grau de correlação linear entre dois conjuntos de variáveis $(\{x_i\} \in \{y_i\})$ e pode ser definido da seguinte forma:

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 (y_i - \overline{y})^2}},$$
(3.1)

em que $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ são, respectivamente, os valores médios de $\{x_i\}$ e $\{y_i\}$. O coeficiente ρ está definido no intervalo $-1 \le \rho \le 1$, sendo que $\rho = 1$ significa uma correlação positiva perfeita, $\rho = 0$ indica que as variavéis não dependem linearmente uma da outra e $\rho = -1$ representa uma correlação negativa perfeita. Para nossos dados, obtemos $\rho \approx 0.05$ para a relação entre a entropia de permutação H e a dimensão típica da imagem. Similarmente, encontramos $\rho \approx 0.01$ para a relação entre a complexidade de permutação C e a dimensão típica da imagem. Esses valores bem pequenos de ρ indicam que praticamente não há dependência linear significativa entre H ou C e o tamanho das imagens.

3.2 Distinguindo os estilos artísticos no plano complexidade-entropia

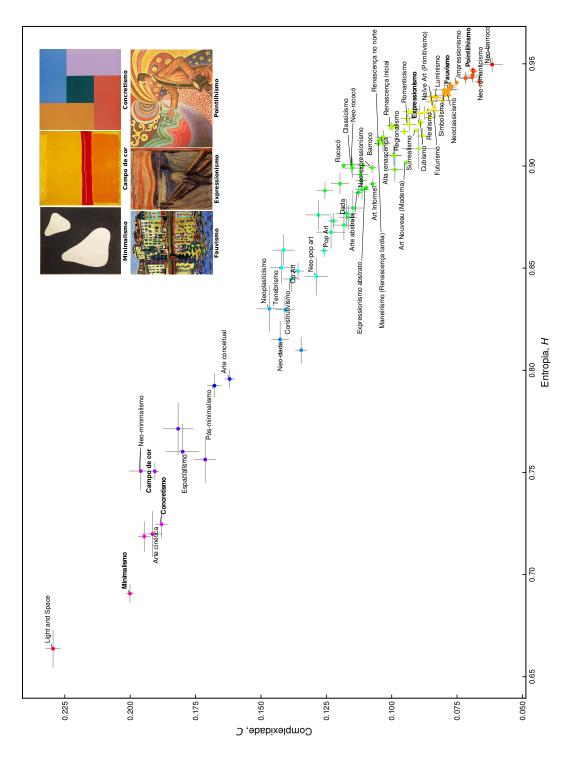
Após verificarmos que as medidas de complexidade estatística não dependem do tamanho das imagens, podemos fazer uso dos valores da entropia H e da complexidade C para realizarmos nossas análises. A primeira pergunta que buscamos responder é se os valores de H e C são capazes de diferenciar os diversos estilos artísticos. Selecionamos os 96 estilos que possuem mais de 100 obras e calculamos o valor médio da entropia e da complexidade de permutação para cada estilo, construindo o plano complexidade-entropia mostrado na figura 3.2. Nesse gráfico, nomeamos apenas alguns estilos para facilitar a leitura.

De maneira geral, observamos que cada estilo artístico ocupa uma posição distinta no plano complexidade-entropia. Além disso, ao inspecionarmos visualmente algumas pinturas desses estilos, é possível verificar que os valores de H e C capturam aspectos relacionados às regularidades ou aleatoriedades dessas obras. Por exemplo, ao analisarmos pinturas de movimentos vanguardistas, como Fauvismo e Expressionismo, em comparação com pinturas dos estilos Minimalista e Concretista, fica evidente essa noção de ordem e regularidade, como ilustram algumas obras de arte mostradas na figura 3.2. Assim, vemos que a entropia de permutação captura o grau de ordem/regularidade da disposição dos pixels das imagens de obras de arte. Estilos artísticos que visualmente apresentam mais regularidades são caracterizados por valores médios menores de entropia. Já aqueles estilos com menor regularidade,

apresentam valores médios maiores de entropia. A complexidade fornece uma informação complementar, na qual estilos com valores maiores de C, em geral, apresentam estruturas regulares maiores (como grandes regiões nas quais existem apenas uma cor). Esses valores altos para a complexidade podem ser comparados com os processos caóticos que analisamos na versão unidimensional do plano complexidade-entropia. Tal como ocorre com sistemas caóticos, pinturas com valores altos de C podem ser caracterizadas pela ausência (ou baixa probabilidade de ocorrência) de certos padrões ordinais. Vale ainda observar que os valores de C são fundamentais para a melhor distinção dos estilos artísticos, visto que existem vários estilos que apresentam valores similares de H e só são devidamente distinguíveis ao considerar a complexidade C.

 \dot{E} interessante mencionar ainda que estilos com algum grau de similaridade estética parecem se agrupar nesse plano. Por exemplo, estilos mais antigos como Barroco, Rococó e a Renascença e suas vertentes (inicial, alta, tardia e no norte) parecem formar um grupo coeso de pontos no plano complexidade-entropia. Já os movimentos vanguardistas (como Cubismo, Expressionismo, Surrealismo, Fauvismo e Simbolismo), parecem formar um outro grupo, com valor médio de entropia maior. Por fim, com os menores valores médios de entropia, notamos um agrupamento dos estilos mais contemporâneos, como Minimalismo, Concretismo, Arte cinética e Neo-minimalismo. Desse modo, além de poder associar a noção mais intuitiva de entropia e complexidade às características de cada estilo artístico, os valores de H e C parecem capturar o grau de similaridade entre estilos que compartilham características estéticas.

Na próxima seção, utilizamos um método de agrupamento hierárquico, baseado nas distâncias entre cada estilo no plano complexidade-entropia, a fim de quantificar de uma maneira mais sistemática esses agrupamentos que são visualmente observados na figura 3.2.



e horizontais) em torno de cada ponto representam os erros padrões das médias de H e C. Vale mencionar que as medidas de e da complexidade C de permutação para cada estilo. Nomeamos apenas alguns para facilitar a leitura. As barras cinzas (verticais complexidade capturam aspectos relacionados às regularidades ou aleatoriedades das obras de arte. As inserções mostram pinturas **Figura 3.2:** Plano complexidade-entropia para os estilos que possuem mais de 100 obras considerando os valores médios da entropia H de alguns estilos artísticos (ver discussão no texto).

3.3 Agrupando os estilos artísticos utilizando um método de agrupamento hierárquico

Na tentativa de identificar os possíveis grupos de estilos artísticos de uma maneira mais sistemática, utilizamos um método de agrupamento hierárquico [40]. Para escolher quais estilos vão se agrupar, uma medida de dissimilaridade e um critério que seja função dessa medida são necessários. Essa medida, geralmente uma métrica com propriedades de distância, quantifica o quão similar um estilo é de outro. De maneira intuitiva, podemos considerar que estilos a uma distância menor são, de alguma forma, mais semelhantes uns aos outros do que os que possuem distância maior. Em nossas análises, utilizamos a distância euclidiana no plano complexidade-entropia como medida de dissimilaridade. A figura 3.3 mostra a matriz das distâncias entre todos os pares de estilos artísticos com mais de 100 obras. Nessa figura, o código de cores indica o valor da distância euclidiana entre cada par de pontos no plano complexidade-entropia, quanto mais escuro o tom de azul, maior é a distância em questão. Analisando essa figura, talvez a característica mais notável seja a maior distância entre os movimentos artísticos mais contemporâneos (como Minimalismo, Arte cinética e Light and Space) e os demais, como também pode ser observado na figura 3.2. Essa característica é manifestada pela existência de linhas (ou colunas) escuras na matriz de distância. A partir da matriz de distância, aplicamos o método de Ward [41] como critério de agrupamento. Esse método, consiste em encontrar os grupos que minimizam a variância da medida de dissimilaridade dentro de cada aglomerado. Em geral, essa busca é feita numericamente utilizando o algoritmo cadeia de primeiros vizinhos [42]. Em nosso caso, utilizamos a função cluster. hierarchy. linkage do módulo de funções Scipy [43], a qual implementa esse procedimento.

O resultado do procedimento de Ward é representado por um dendrograma (um diagrama em forma de árvore), como o mostrado na figura 3.4. Nesse gráfico, as barras horizontais indicam a distância entre dois estilos de pinturas que estão conectados por barras verticais. No caso da conexão ser com um grupo de estilos, o comprimento da barra vertical indica a menor das distâncias aos elementos do grupo.

Por construção, um dendrograma mostra como é a estrutura hierárquica associada à matriz de distância, revelando os grupos com as menores distâncias entre si. No caso dos estilos artísticos, observamos a presença de alguns agrupamentos de movimentos artísticos de períodos similares e/ou com características estéticas parecidas. Como exemplo, vamos analisar algumas das conexões dentro de um desses grupos. Notamos que os estilos mais antigos, como Barroco e as vertentes da Renascença estão agrupados. Nesse mesmo grupo, o Realismo Fantástico também está presente, um movimento que emergiu no fim dos anos 1950 e, como o próprio nome sugere, representa a fantasia, imaginação e mitologia. Essa também é uma característica particularmente importante para o Maneirismo, um movimento

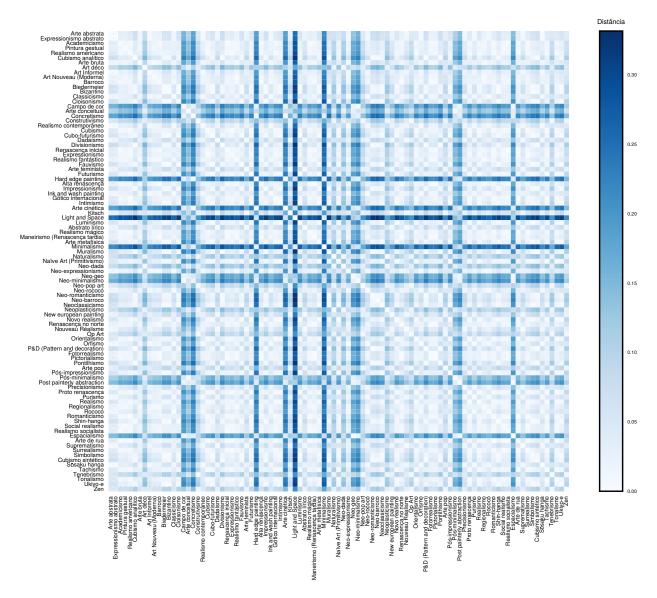


Figura 3.3: Matriz da distância euclidiana no plano complexidade-entropia entre todos os pares de estilos artísticos com mais de 100 obras. O código de cores indica o valor dessa distância. Quanto menor a distância, mais claro é o tom de azul e, portanto, maior o grau de similaridade entre os estilos.

da Renascença tardia que surgiu por volta de 1600. Percebemos que, mesmo havendo uma disparidade temporal enorme, as características estéticas semelhantes são capturadas. Outro exemplo desse mesmo grupo é o Pictorialismo e o Fotorrealismo. O primeiro é considerado o movimento que deu origem a fotografia moderna, em que os artistas tentavam por meio de suas imagens fotográficas fazer uma aproximação à pintura, alterando a granulação, os tons e modificando elementos de forma a assemelhar fotografias às pinturas. Os pictorialistas eram os amantes de fotografias que tentavam reproduzir ideias da pintura, da arte barroca e clássica, e assim elevar a importância da fotografia ao *status* de arte. Já no Fotorrealismo, os artistas buscam representar com bastante fidelidade (em suas pinturas) imagens origi-

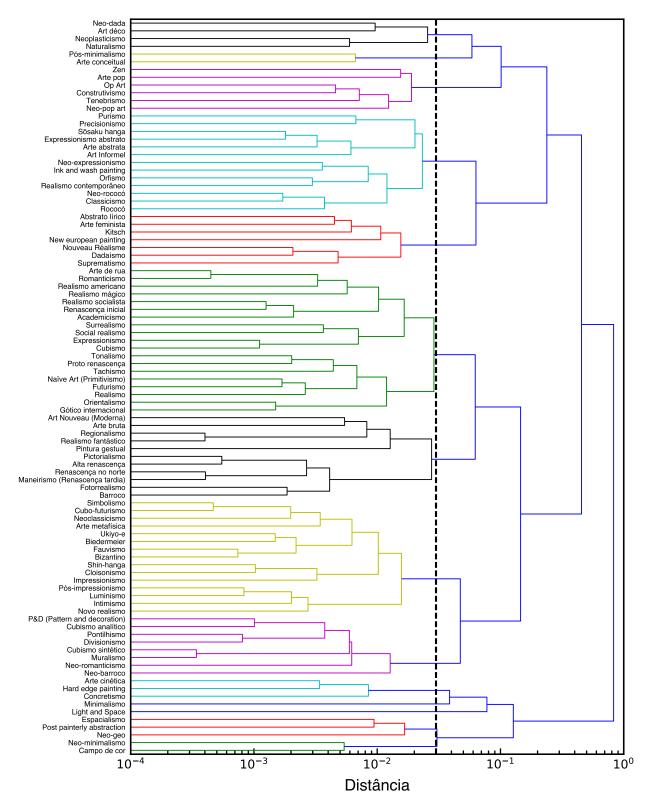


Figura 3.4: Dendrograma evidenciando os grupos que foram formados após realizarmos o agrupamento usando a distância limiar que maximiza o coeficiente de silhueta. A linha tracejada indica o valor dessa distância (0,03).

nalmente obtidas com uma câmera fotográfica. Assim, no Pictorialismo, as fotografias se parecem com pinturas e, no Fotorrealismo, as pinturas que se parecem com fotografias.

De uma maneira geral, observamos que podemos definir uma distância limiar para produzir agrupamentos a partir do dendrograma da figura 3.4. Por exemplo, escolhendo essa distância igual a 0,3, observamos que o dendrograma é dividido em três grandes grupos de estilos. Valores menores para essa distância limiar tendem a produzir um maior número de grupos. Nesse contexto, a pergunta mais natural é como escolher um valor para essa distância limiar. A resposta para essa pergunta não é única, mas uma possível maneira para escolher esse valor limiar é considerar o chamado coeficiente de silhueta [44]. Esse coeficiente quantifica a qualidade do agrupamento e é definido para cada estilo *i* como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$
(3.2)

sendo a(i) a dissimilaridade média do estilo i dentro de seu grupo (também chamada de coesão, no nosso caso, o valor médio das distâncias que separam o estilo i dos demais estilos de seu grupo), b(i) a dissimilaridade média do estilo i com os estilos do grupo mais próximo (também chamada de separação) e $\max\{a(i),b(i)\}$ o maior valor entre essas duas quantidades.

Sendo assim, o coeficiente de silhueta global $\langle s \rangle$ é o valor médio de s(i) sobre todos os estilos. Do modo como é definido, o coeficiente de silhueta varia entre -1 e 1, sendo que quanto mais próximo de 1 ele for, melhor é a qualidade do agrupamento. De maneira intuitiva, para $\langle s \rangle \approx 1$ é necessário que $a(i) \ll b(i)$ para todos os estilos, ou seja, como a(i) mede a dissimilaridade dentro de seu próprio grupo, um valor pequeno de a(i) significa que o estilo i é bem próximo dos demais de seu grupo e, então, foi bem agrupado. De maneira similar, $\langle s \rangle \approx -1$ se $b(i) \ll a(i)$ para todos os estilos, 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 divide o dendrograma é procurar pelo valor dessa distância que maximiza o coeficiente global de silhueta. Essa análise é mostrada na figura 3.5, na qual observamos o valor desse coeficiente em função da distância limiar. Notamos 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 dividir o dendrograma da figura 3.4. De fato, nessa figura, a linha tracejada indica esse limiar e as diferentes cores do dendrograma mostram os agrupamentos de estilos artísticos obtidos. Os 96 estilos com mais de 100 obras de arte formam 14 grupos, conforme também é mostrado na tabela 3.1. O maior desses grupos contêm 19 estilos artísticos, além disso, existem dois estilos que não são agrupados com nenhum outro (Minimalismo e Light and Space). Alguns desses grupos parecem fazer sentido histórico e estético, como o grupo 12

(que analisamos anteriormente) e o agrupamento de movimentos vanguardistas (grupo 5), como Cubismo, Futurismo e Expressionismo e o Realismo e suas vertentes. Outros parecem ser mais complicados de explicar apenas pela história da arte, como o grupo 1, em que o Naturalismo está presente, um estilo do período Renascentista, no qual os artistas representam objetos realistas com atenção a cada detalhe, e o Neoplasticismo, um movimento de vanguarda do início do século XX totalmente relacionado à arte abstrata. Entretanto, e apesar de nosso conhecimento limitado sobre arte, podemos afirmar que a estrutura hierárquica produzida a partir do plano complexidade-entropia faz sentido de maneira geral, podendo, eventualmente, ser melhor descrita por especialistas em arte e história da arte.

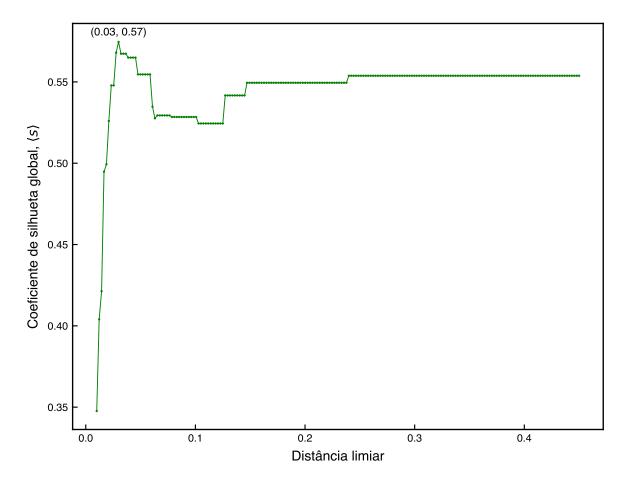


Figura 3.5: Coeficiente de silhueta global $\langle s \rangle$ levando em conta todos os estilos artísticos que possuem mais de 100 obras. O valor máximo desse coeficiente e, consequentemente, o que maximiza a qualidade do agrupamento, ocorre para a distância de 0,03, que foi utilizada como limiar para nossa análise.

Por fim, na figura 3.6 mostramos uma visualização combinada entre a matriz de distância e o dendrograma, costumeiramente denominada de *clustermap*. Essa figura mostra a matriz de distância numa ordem definida pelos agrupamentos obtidos por meio do dendrograma e deve ser encarada como um resumo da análise de agrupamentos hierárquicos dos estilos artísticos de pintura em nossa base de dados.

Tabela 3.1: Os 14 grupos de estilos artísticos formados após escolhermos o valor 0,03 como distância limiar para dividir o dendrograma da figura 3.4.

	Neo-dada		Pós-minimalismo
	Art déco	Grupo 6	Arte conceitual
Grupo 1	Neoplasticismo		Neo-minimalismo
	Naturalismo	Grupo 7	Campo de cor
	Abstrato lírico		Zen
	Arte feminista	Grupo 8	Arte pop
Grupo 2	Kitsch		Op art
	New european painting		Construtivismo
	Nouveau Réalisme		Tenebrismo
	Dadaísmo		Neo-pop art
	Suprematismo	Grupo 9	Arte cinética
	P & D (Pattern and decoration)		Hard edge painting
	Cubismo analítico		Concretismo
	Pontilhismo		Simbolismo
	Divisionismo		Cubo-futurismo
Grupo 3	Cubismo sintético		Neoclassicismo
	Muralismo		Arte metafísica
	Neo-romanticismo		
			Ukiyo-e
	Neo-barroco		Biedermeier
	Purismo	G 10	Fauvismo
	Precisionismo	Grupo 10	Bizantino
	Sōsaku hanga		Shin-hanga
	Expressionismo abstrato		Cloisonismo
	Arte abstrata		Impressionismo
~ .	Art Informel		Pós-impressionismo
Grupo 4	Neo-expressionismo		Luminismo
	Ink and wash painting		Intimismo
	Orfismo		Novo realismo
	Realismo contemporâneo	Grupo 11	Espacialismo
	Neo-rococó		Post painterly abstraction
	Classicismo		Neo-geo
	Rococó		Arte Nouveau (Moderna)
	Arte de rua		Arte bruta
	Romanticismo		Regionalismo
	Realismo americano		Realismo fantástico
	Realismo mágico		Pintura gestual
	Realismo socialista	Grupo 12	Pictorialismo
	Renascença inicial		Alta renascença
	Academicismo		Renascença no norte
	Surrealismo		Maneirismo (Renascença tardia)
	Social realismo		Fotorrealismo
Grupo 5	Expressionismo		Barroco
	Cubismo	Grupo 13	Minimalismo
	Tonalismo	Grupo 14	Light and Space
	Proto renascença		
	Tachismo		
	Naïve Art (Primitivismo)		
	Futurismo		
	Realismo		
	Orientalismo		
	Gótico internacional		
	·	1	

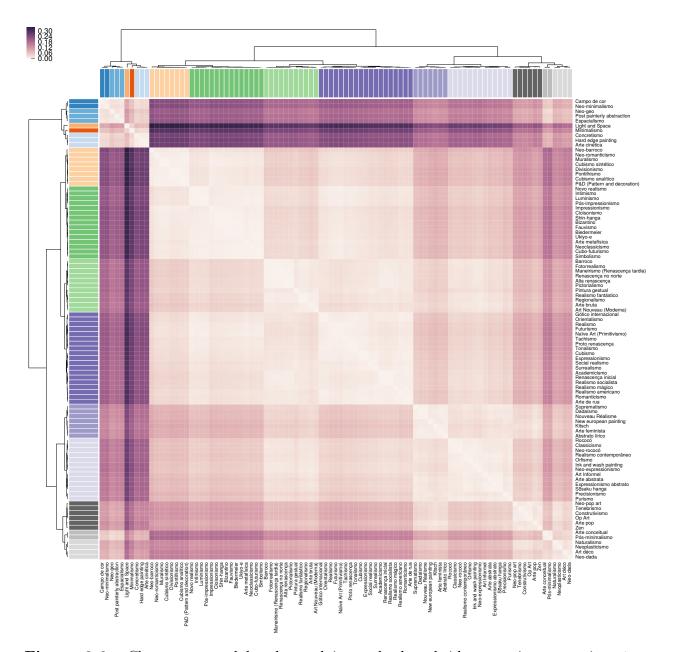


Figura 3.6: Clustermap englobando os dois resultados obtidos anteriormente, isto é, o dendrograma e a matriz de distância numa ordem compatível com os agrupamentos dos estilos artísticos.

3.4 Evolução temporal das medidas de complexidade estatística no plano complexidade-entropia

Como uma última análise a ser apresentada nesse trabalho, estudamos a evolução temporal de H e C no plano complexidade-entropia. Essa evolução é analisada de forma geral, ou seja, não selecionamos por estilos artísticos específicos. Na figura 3.7, mostramos a evolução temporal das medidas de complexidade estatística H e C no plano complexidade-entropia, levando em conta os valores médios dessas grandezas dentro das janelas de tempo indicadas na figura. Notamos que os intervalos de tempo não possuem o mesmo tamanho. Isso ocorre pois, para que a média seja representativa, dividimos a base de dados em 20 intervalos de tempo, de modo que em cada um contenha (aproximadamente) o mesmo número de obras de arte. Essa escolha se deve ao fato da grande maioria das obras de nossa base de dados ter sido produzida a partir de 1850, como já apresentamos na figura 2.1. Por isso, os três primeiros pontos da figura 3.7 possuem um intervalo de tempo bem maior do que os outros.

Analisando a figura 3.7, notamos uma evolução temporal clara ao longo da história da arte e do desenvolvimento dos movimentos artísticos. Buscamos por acontecimentos marcantes que possam, talvez, ser associados à evolução observada. Primeiramente, destacamos o fato das principais divisões da história da arte em períodos serem claramente observadas em nossa análise. Especificamente, o período da Arte Moderna é compreendido entre 1870 (com o surgimento do Impressionismo na França e, posteriormente, com o desenvolvimento dos movimentos vanguardistas, durante as décadas iniciais do século XX) até 1970, quando se dá o ínicio do período da chamada Arte Pós-Modernista, considerada como uma ampla categoria de arte contemporânea criada de 1970 em diante [46]. Essa divisão é evidenciada pelas áreas sombreadas na figura 3.7. Notamos que, pelo menos durante as décadas iniciais do período correspondente à Arte Moderna (que também foram as mais importantes com relação aos movimentos que nelas se desenvolveram, como mostrado na linha do tempo da figura 3.8), os valores tanto de entropia e de complexidade permanecem localizados em torno de um pequeno intervalo. Intrigantemente, essa mesma característica é observada para o período correspondente à Arte Pós-Modernista.

Percebemos que, com o passar do tempo, há um deslocamento claro para os valores menores de entropia e maiores de complexidade, concordando com o que já observamos ao analisar a localização individual dos estilos artísticos no plano complexidade-entropia, mostrada na figura 3.2. Assim, ao confrontarmos essas duas figuras, observamos que as características estéticas (em geral) refletem àquelas dos estilos predominantes em cada época.

Finalmente, tentamos analisar de maneira mais detalhada o período no qual essa transição ocorre. Ao inspecionarmos a figura 3.7, notamos que essa transição se inicia entre os anos 1909 e 1914. Em uma pesquisa por fatos marcantes desse período, nos deparamos inúmeras vezes com o nome de Pablo Picasso, um dos mais renomados artistas plásticos do século XX

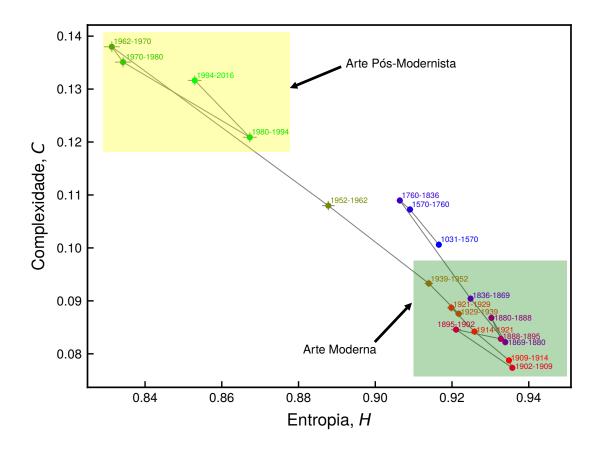
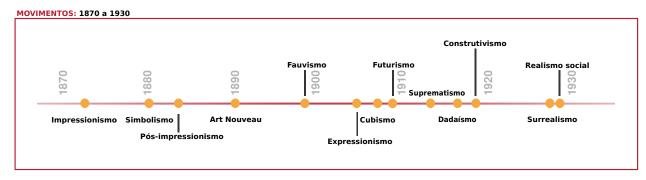


Figura 3.7: Evolução temporal de H e C no plano complexidade-entropia levando em conta os valores médios dessas grandezas dentro de cada janela de tempo (indicada no gráfico). Os intervalos de tempo foram escolhidos dividindo as pinturas em 20 grupos com (aproximadamente) o mesmo número de obras de arte. As áreas sombreadas contêm os períodos correspondentes à Arte Moderna e à Arte Pós-Modernista. Notamos que com o passar do tempo há um deslocamento para os valores menores de entropia e maiores de complexidade.

que viveu durante os anos 1881 e 1973. Estudando mais especificamente sua contribuição no desenvolvimento da arte, é possível assumir que Picasso seja, talvez, o pintor mais proeminente de sua época. Como uma observação importante, destacamos o fato de no ano 1907 ele ter criado uma de suas obras-primas, Les Demoiselles d'Avignon, considerada um marco no desenvolvimento da Arte Moderna [46]. Essa obra é associada à primeira quebra radical das ideias artísticas dos períodos anteriores e fundamental para a criação do Cubismo, considerado um dos movimentos artísticos mais influentes do século XX [46,47]. De fato, Pablo Picasso juntamente com George Braque são considerados os fundadores desse movimento artístico revolucionário, que se desenvolveu por volta de 1908, e, entre outras características, se contrapõe às ideias convencionais de perspectiva e forma. Posteriormente, com a Primeira Guerra Mundial (1914) e a Revolução Russa (1917), muitas ideias convencionais do mundo



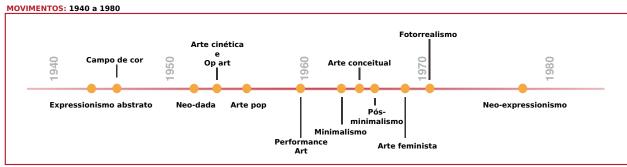


Figura 3.8: Linha do tempo mostrando alguns dos movimentos artísticos de cada período. Vale notar que foi durante as décadas iniciais do período da Arte Moderna (que se inicia com o Impressionismo em 1870) que se desenvolveram a maioria dos movimentos artísticos mais importantes daquela época. Figura adaptada da referência [45].

da pintura e da escultura são modificadas, levando ao surgimento de inúmeros movimentos vanguardistas [46].

A despeito de nosso conhecimento muito limitado sobre história da arte, acreditamos que, de maneira geral, a evolução temporal observada no plano complexidade-entropia parece fazer sentido com alguns fatos marcantes no desenvolvimento das artes visuais.

Conclusões e perspectivas

Nesse trabalho, estudamos aspectos das artes visuais utilizando ferramentas da Física Estatística. Fizemos uso de apenas duas medidas de complexidade estatística (a entropia H e a complexidade C de permutação) aplicadas às imagens de uma das maiores bases de dados online sobre artes visuais chamada $Wiki\ Art$. Dessa base, utilizamos aproximadamente 100 mil imagens e vários metadados, como estilo artístico, data e gênero das obras de arte. Para fazer o download e tratar essa grande quantidade de dados, utilizamos pacotes e programas escritos em linguagem de programação Python, tanto para extrair os dados de forma automatizada, quanto para visualizar os resultados das análises.

É importante ressaltarmos o relativo sucesso obtido ao utilizar o plano complexidadeentropia para diferenciar os diversos estilos artísticos, capturando as características relacionadas às regularidades e/ou aleatoriedades das imagens. Posteriormente, utilizamos um método de agrupamento hierárquico para encontrar os grupos de estilos que são visualmente observamos nesse plano. Apesar de nosso pouco conhecimento a respeito da história da arte, constatamos que, de forma geral, os agrupamentos fazem sentido histórico e estético.

Em uma última análise, buscamos quantificar a evolução temporal dos movimentos artísticos ao longo do desenvolvimento das artes visuais. Observamos uma evolução clara dos valores de H e C no plano complexidade-entropia, na qual há um deslocamento para valores menores de entropia e maiores de complexidade com o passar do tempo. Essa evolução observada vai de encontro com o resultado obtido ao classificarmos os estilos nesse plano. Além disso, a divisão da história da arte em períodos, especificamente Arte Moderna e Arte Pós-Modernista, também é capturada em nossa análise, ou seja, percebemos que as características estéticas (em geral) refletem àquelas dos estilos predominantes em cada época. Esses fatos podem motivar novas análises e métodos que possam ser aplicados nesse campo, o qual é pouco estudado de forma quantitativa.

Por fim, como perspectivas de futuros estudos, podemos citar o estudo de artistas es-

pecíficos (os mais renomados, por exemplo) e, assim, procurar entender as características estéticas e evolutivas de pintores de maneira quantitativa. Uma outra possibilidade é construir uma rede complexa das pinturas parecidas para tentar identificar (de modo imparcial) as obras e pintores mais importantes e influentes. Numa direção mais aplicada, acreditamos que nosso estudo possa contribuir para a construção de aplicativos para a classificação automatizada de pinturas, assim como já existem para músicas. Ressaltamos que existem muitas possibilidades para serem exploradas no estudo quantitativo de artes visuais, de modo que nosso trabalho representa uma pequena contribuição na tentativa de preencher essa lacuna e contribuir para o avanço das aplicações interdisciplinares da Física.

Referências Bibliográficas

- [1] Boccara, N. Modeling complex systems (Springer-Verlag, New York, 2004).
- [2] de Picoli Junior, S. Física estatística dos sistemas complexos: aplicações interdisciplinares. *Universidade Estadual de Maringá Tese de Doutorado* (2007).
- [3] Ribeiro, H. V. Identificação e modelagem de padrões em sistemas complexos. *Universidade Estadual de Maringá Tese de Doutorado* (2012).
- [4] Antonio, F. J. Sistemas complexos: Desastre, loteria e epidemias. *Universidade Estadual de Maringá Tese de Doutorado* (2013).
- [5] Tateishi, A. A. Processos estocásticos e difusão anômala em sistemas complexos. *Universidade Estadual de Maringá Tese de Doutorado* (2013).
- [6] dos Santos, E. S. Controle postural como um sistema complexo: Análise da distribuição das velocidades do centro-de-pressão. *Universidade Estadual de Maringá Tese de Doutorado* (2015).
- [7] Mantovani, M. C. Física estatística aplicada a processos eletoriais. *Universidade Estadual de Maringá Tese de Doutorado* (2015).
- [8] da Silva, A. I. Atividade psicomotora , epidemias e lideranças como sistemas complexos. Universidade Estadual de Maringá - Tese de Doutorado (2015).
- [9] Guimarães, R. R. Aplicação de dinâmica difusiva: Peixes e portadores de carga. *Universidade Estadual de Maringá Tese de Doutorado* (2016).
- [10] de Andrade Alves, L. G. Invariância de escala e difusão anômala em sistemas complexos urbanos e biológicos. *Universidade Estadual de Maringá Tese de Doutorado* (2017).

- [11] Bandt, C. & Pompe, B. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters* **88**, 174102 (2002).
- [12] López-Ruiz, R., Mancini, H. L. & Calbet, X. A statistical measure of complexity. *Physics Letters A* **209**, 321–326 (1995).
- [13] Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A. & Fuentes, M. A. Distinguishing noise from chaos. *Physical Review Letters* **99**, 154102 (2007).
- [14] 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).
- [15] Anônimo. Reality check on reproducibility. Nature News 533, 437 (2016).
- [16] Ward, A., Baldwin, T. O. & Antin, P. B. Research data: Silver lining to irreproducibility. Nature 532, 177–177 (2016).
- [17] Anônimo. Repetitive flaws. Nature News 529, 256 (2016).
- [18] Bissell, M. Reproducibility: The risks of the replication drive. *Nature News* **503**, 333 (2013).
- [19] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423 (1948).
- [20] Courant, R. Differential and Integral Calculus, Vol. 2 (Wiley-Interscience, New York, 1988).
- [21] Rabiner, L. R. & Schafer, R. W. Digital Processing of Speech Signals (Prentice-Hall, 1978).
- [22] Crandall, I. B. & MacKenzie, D. Analysis of the energy distribution in speech. *Bell System Technical Journal* 1, 116–128 (1922).
- [23] Martin, M. T., Plastino, A. & Rosso, O. A. Statistical complexity and disequilibrium. *Physics Letters A* **311**, 126–132 (2003).
- [24] 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).
- [25] Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52**, 479–487 (1988).
- [26] Rényi, A. On measures of entropy and information (The Regents of the University of California, 1961).

- [27] Wootters, W. K. Statistical distance and Hilbert space. *Physical Review D* **23**, 357–362 (1981).
- [28] Grosse, I. et al. Analysis of symbolic sequences using the Jensen-Shannon divergence. Physical Review E 65, 041905 (2002).
- [29] Brito, A. F., Redinz, J. A. & Plascak, J. A. Dynamics of rough surfaces generated by two-dimensional lattice spin models. *Physical Review E* **75**, 046106 (2007).
- [30] Tomé, T. & de Oliveira, M. J. Dinâmica Estocástica e Irreversibilidade (EdUSP, 2001).
- [31] Lyu, S., Rockmore, D. & Farid, H. A digital technique for art authentication. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 17006–17010 (2004).
- [32] Johnson, C. et al. Image processing for artist identification. *IEEE Signal Processing Magazine* **25**, 37–48 (2008).
- [33] Coddington, J., Elton, J., Rockmore, D. & Wang, Y. Multifractal analysis and authentication of Jackson Pollock paintings. In *Electronic Imaging 2008*, 68100F–68100F (International Society for Optics and Photonics, 2008).
- [34] 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* **107**, 1279–1283 (2010).
- [35] Walker, F., Bucker, B., Anderson, N. C., Schreij, D. & Theeuwes, J. Looking at paintings in the Vincent Van Gogh Museum: Eye movement patterns of children and adults. *PLOS ONE* 12, e0178912 (2017).
- [36] Python.org. URL https://www.python.org/.
- [37] Scrapy | a fast and powerful scraping and web crawling framework. URL https://scrapy.org/.
- [38] OpenCV: OpenCV modules. URL http://docs.opencv.org/trunk/index.html.
- [39] Hogg, R. V., McKean, J. & Craig, A. T. Introduction to Mathematical Statistics (Pearson, Boston, 2012).
- [40] Maimon, O. & Rokach, L. *Data Mining and Knowledge Discovery Handbook* (Springer Science & Business Media, 2010).

- [41] Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236–244 (1963).
- [42] Müllner, D. Modern hierarchical, agglomerative clustering algorithms. arXiv:1109.2378 (2011).
- [43] SciPy.org. URL https://scipy.org/.
- [44] 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).
- [45] The art story: Modern art movement timeline. URL http://www.theartstory.org/section_movements_timeline.htm.
- [46] History of art timeline. URL http://www.visual-arts-cork.com/history-of-art-timeline.htm.
- [47] Cubism: The Leonard A. Lauder Collection | The Metropolitan Museum of Art. URL http://www.metmuseum.org/press/exhibitions/2014/cubism-the-leonard-lauder-collection.