UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Leonardo Gabriel José Mendes Voltarelli

Entropia de permutação dos vizinhos mais próximos

Maringá, 30 de setembro de 2024.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Leonardo Gabriel José Mendes Voltarelli

Entropia de permutação dos vizinhos mais próximos

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, 30 de setembro de 2024.

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

Voltarelli, Leonardo Gabriel José Mendes

V935e Entropia de permutação dos vizinhos mais próximos / Leonardo Gabriel José Mendes Voltarelli. -- Maringá, PR, 2024.
70 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, 2024.

1. Entropia. 2. Padrões espaciais. 3. Imagens. 4. Séries temporais. 5. Sistemas complexos. 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. 536.73

Elaine Cristina Soares Lira - CRB-9/1202

LEONARDO GABRIEL JOSÉ MENDES VOLTARELLI

ENTROPIA DE PERMUTAÇÃO DOS VIZINHOS MAIS PRÓXIMOS

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

Aprovado em: Maringá, 27 de setembro de 2024.

BANCA EXAMINADORA

_____

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

_____

Prof. Dr. Thomas Kauê Dal'Maso Peron
Universidade de São Paulo - USP/ICMC

_____

Prof. Dr. Rafael Soares Zola
Universidade Tecnológica Federal do Paraná - UTFPR

Agradecimentos

Primeiramente, gostaria de agradecer a minha família, principalmente a minha mãe Janete e meu irmão Eduardo, que me apoiaram desde o princípio e me motivaram a seguir o caminho das ciências. Também gostaria de dedicar um agradecimento especial à minha avó, que apesar de ter suas raízes em uma realidade rural e nunca ter recebido educação formal, ainda assim entendia a importância dos estudos. Ela sempre se orgulhou muito de eu ter ingressado em uma universidade pública e ter seguido na carreira acadêmica, sendo o primeiro da nossa família a fazê-lo. Estudar sem preocupações financeiras é o resultado do longo esforço da minha avó em proporcionar melhores condições para nossa família. A vida que temos agora é o legado de tudo que ela fez por nós. Embora ela não esteja mais conosco, sua influência continua presente em todos os aspectos da minha vida. O sonho dela de uma vida melhor é o que me motiva hoje.

Agradeço imensamente ao Professor Haroldo Valentin Ribeiro pela paciência e disposição em me orientar durante o mestrado. Também sou grato ao Professor Renio, que, atendendo a um pedido meu, utilizou o livro de Landau na disciplina de Eletrodinâmica, o que fez dessa matéria a mais bonita que já vi. Agradeço aos amigos do ComplexLab que me ajudaram desde o início, proporcionando um ambiente agradável e descontraído, com o equilíbrio perfeito entre leveza e seriedade nos momentos certos.

Agradeço à minha namorada, Mayra, pelo seu amor, paciência e apoio incondicional durante todo este processo.

Agradeço à CAPES e ao CNPq pelo auxílio financeiro; sem ele seria impossível ter realizado esse trabalho.

Resumo

A entropia de permutação e seus métodos derivados são técnicas inspiradas na Física e eficazes no processamento de dados complexos. Contudo, seu uso tem sido limitado a dados estruturados, como séries temporais ou imagens. Nesta dissertação, introduzimos a entropia de permutação dos k-primeiros vizinhos, uma extensão inovadora adaptada para dados não estruturados, independentemente de sua configuração espacial, temporal ou dimensionalidade. No Capítulo 1, apresentamos essa nova abordagem, composta de três etapas fundamentais: primeiro, cria-se um grafo a partir da conexão dos primeiros vizinhos de cada ponto dos dados. Em seguida, por meio de caminhadas aleatórias dentro desse grafo, extraem-se séries temporais dos valores dos pontos. Finalmente, utiliza-se a representação simbólica para obter uma distribuição de padrões ordinais dessas séries, o que possibilita calcular a entropia. No Capítulo 2, apresentamos experimentos computacionais que comprovam a eficácia da nossa medida em distinguir entre regimes regulares e aleatórios em dados dispersos, tanto no tempo quanto no espaço. Utilizamos o movimento browniano fracionário como modelo e comparamos nossa medida com o índice de Moran, demonstrando desempenho superior quando avaliada por métodos de aprendizado de máquina. Também aplicamos nossa técnica na distinção entre assinaturas genuínas e falsificadas, utilizando a distribuição ordinal. Em comparação com a entropia de permutação original, nossa métrica mostrou-se superior na verificação da autenticidade das assinaturas. O Capítulo 3 desenvolve uma extensão do método para cálculo de nossa entropia em imagens, tratando os pixels como dados distribuídos em uma rede regular. Para verificar essa abordagem, utilizamos nossa medida para prever o passo de texturas de cristais líquidos colestéricos e comparamos os resultados com outra abordagem que utiliza a entropia de permutação usual, constatando um desempenho superior de nossa técnica. No Capítulo 4, apresentamos a extensão do método para séries temporais não regulares no tempo. Tratamos as séries como dados espalhados em duas dimensões, oferecendo uma abordagem natural para incorporar informações de amplitude e lacunas temporais, melhorando significativamente a resiliência ao ruído e as capacidades preditivas em comparação com a entropia de permutação usual.

Palavras-chave: Entropia. Padrões Espaciais. Imagens. Séries Temporais. Sistemas Complexos. Ciência de Dados.

Abstract

Permutation entropy and its derived analysis methods are physics-inspired techniques effective in processing complex and extensive datasets. However, despite substantial progress in developing and applying these tools, their use has predominantly been limited to structured datasets such as time series or images. In this dissertation, we introduce k-nearest neighbor permutation entropy, an innovative extension adapted for unstructured data, regardless of its spatial, temporal, or dimensional configuration. Chapter 1 presents this new approach, which consists of three fundamental steps: first, a graph is created by connecting the nearest neighbors of each data point; then, through random walks within this graph, time series of the data point values are extracted; finally, a symbolic representation is used to obtain a distribution of ordinal patterns from these series, enabling the calculation of entropy. In Chapter 2, we discuss numerical experiments conducted to demonstrate the effectiveness of our entropy measure in distinguishing between regular and random regimes in scattered data, both in time and space. We used fractional Brownian motions as a test case and compared our measure with Moran's index. Our approach showed significantly superior performance when evaluated using machine learning methods. Additionally, we applied our technique to distinguish between genuine and forged signatures using the ordinal distribution. Compared to a similar method that uses the usual permutation entropy, our approach proved superior in the task of verifying signature authenticity. Chapter 3 develops an extension of our method that allows the calculation of our entropy in images by treating pixels as data distributed on a regular grid. To verify the capacity of this approach, we used our metric to predict the pitch of cholesteric liquid crystal textures and compared the results with another approach that uses the usual permutation entropy. We found that our measure performs better. In Chapter 4, we present the extension of our method for application in time series that are not regular in time. To do this, we treat the series as data scattered in two dimensions, offering a natural approach to incorporating amplitude information and temporal gaps, significantly improving noise resilience and predictive capabilities compared to usual permutation entropy.

Keywords: Entropy. Spatial Patterns. Images. Time Series. Complex Systems. Data Science.

Sumário

Introdução

Os campos científicos estão testemunhando universalmente um aumento sem precedentes no volume e na complexidade dos dados digitais disponíveis para pesquisa. Esta revolução dos dados [1] estimulou uma demanda urgente por métodos que sejam simultaneamente simples, interpretáveis, robustos, computacionalmente eficientes e, ainda assim, capazes de extrair informações valiosas de grandes bases de dados. Aproveitando uma longa tradição de descobrir princípios fundamentais a partir de sistemas complexos, métodos inspirados na física emergiram como notavelmente eficazes para enfrentar esses desafios. Um exemplo relevante é a entropia de permutação [2] e sua ideia subjacente de que a ordem relativa dos pontos de dados é instrumental para a caracterização do sistema. Este método não apenas atende aos requisitos para gerenciar dados complexos e volumosos, mas também oferece o benefício adicional de facilitar resultados reprodutíveis. A entropia de permutação encontrou aplicações bem-sucedidas relacionadas à análise de dados em diversas disciplinas. No campo da engenharia, foi possível detectar e amplificar efetivamente as mudanças dinâmicas de sinais de vibração gerados por máquinas rotativas (dispositivos mecânicos que utilizam movimento rotacional para realizar trabalho) e caracterizar seus estados de operação sob diferentes condições de operação [3]. Nas ciências biomédicas, a entropia de permutação foi utilizada para gerar um detector automatizado de crises epilépticas [4]. Em econofísica, ela pode ser utilizada para quantificar o estágio de desenvolvimento do mercado de ações e sua ineficiência [5]. Essa medida também se demonstrou útil em ciências climáticas, sendo usada para detectar anomalias em registros de isótopos de água em dados de um núcleo de gelo polar profundo [6]. A entropia de permutação também é sensível a transições de fases em cristais líquidos [7, 8], sendo assim uma ferramenta útil na análise de matéria condensada. Um caso especialmente interessante é que ela pode ser usada até mesmo nas artes visuais [9], evidenciando sua versatilidade e utilidade em diversos domínios de pesquisa [10–14].

O sucesso da entropia de permutação vai além de suas aplicações práticas, pois seu prin-

cípio central de derivar uma distribuição de probabilidade a partir de padrões ordinais em dados serviu como estrutura fundamental para o desenvolvimento de uma infinidade de ferramentas de análise de dados. Esses avanços incluem o cálculo de outros quantificadores a partir da distribuição de probabilidade ordinal, como a entropia condicional [15], o plano complexidade-entropia [16] e as curvas complexidade-entropia [17]. Também abrangem a análise de padrões proibidos [18, 19], o tratamento de padrões de valores iguais [20, 21] e a exploração de padrões ordinais em múltiplas escalas [22]. A metodologia original também inspirou a criação de redes ordinais [23–28], que exploram transições entre padrões ordinais, revelando novas perspectivas sobre a dinâmica de sistemas complexos. Além disso, a metodologia subjacente à entropia de permutação foi estendida para acomodar dados bidimensionais [29–31], ampliando significativamente sua aplicabilidade à análise de conjuntos de dados de imagens.

Apesar do progresso teórico substancial e da aplicação bem-sucedida na análise de dados, os métodos baseados na entropia de permutação são limitados a dados estruturados em grade. Essa restrição decorre da conceitualização inicial da entropia de permutação, que se concentrou em discernir padrões de ordenação entre observações sequenciais em séries temporais [2]. A extensão para dados bidimensionais herda essa limitação, concentrando-se em padrões ordinais derivados de partições retangulares [29]. Além disso, mesmo com generalizações multiescala [22,30], a análise ainda depende de partições em grade compostas por elementos espaçados uniformemente, com a ideia de múltiplas escalas referindo-se ao uso de diferentes intervalos de tempo para amostragem de séries temporais [22] ou ajuste de níveis de resolução na análise de imagens [30]. Superar essa limitação tem o potencial de expandir significativamente a utilidade da entropia de permutação para uma variedade maior de tipos de dados, incluindo dados de nuvem de pontos (como varreduras tridimensionais de sensores LiDAR ou imagens médicas), dados de processos pontuais (como distribuição espacial de surtos de doenças ou ocorrências de terremotos), dados geoespaciais (como rastreamento GPS de animais, indicadores urbanos em sistemas de cidades ou pontos de amostragem relacionados ao monitoramento ambiental) e várias outras estruturas de dados que não são amostradas uniformemente no tempo ou no espaço. Além disso, enfrentar esse desafio abre vias promissoras de pesquisa relacionadas à extensão e adaptação do extenso conjunto de ferramentas derivadas da entropia de permutação para essas várias estruturas de dados.

Nesta dissertação, introduzimos a entropia de permutação dos k-primeiros vizinhos [32], uma generalização inovadora da entropia de permutação projetada para acomodar dados irregulares ou fora de grade em espaços multidimensionais. Nosso método emprega um grafo de vizinhos mais próximos para estabelecer relações de vizinhança entre os pontos de dados e usa caminhadas aleatórias nesse grafo para gerar séries temporais. Essas séries temporais permitem a extração de padrões ordinais, o cálculo de sua distribuição de probabilidade e a estimativa da entropia de Shannon, definindo assim a entropia de permutação dos k-primeiros

vizinhos de estruturas de dados gerais, independentemente de sua configuração espacial ou temporal e dimensionalidade. Investigamos a eficácia dessa nova técnica na análise de padrões espaciais, revelando que ela não só identifica habilmente variações nesses padrões, mas também o faz com um nível de acurácia que supera significativamente medidas convencionais como a autocorrelação espacial. Demonstramos ainda que nossa abordagem melhora a caracterização de séries temporais amostradas irregularmente ao incorporar intuitivamente informações sobre lacunas temporais. Além de ampliar o escopo da abordagem padrão, mostramos que a entropia de permutação dos k-primeiros vizinhos é igualmente aplicável a séries temporais regulares e imagens, mas oferece maior robustez contra ruído e melhor capacidade de discriminação em comparação com a entropia de permutação convencional.

O restante do presente trabalho está estruturado da seguinte maneira. No Capítulo 1, apresentamos os métodos necessários para a obtenção da entropia dos k-primeiros. Em seguida, no Capítulo 2, apresentamos nossos principais resultados, que incluem o cálculo da entropia em processos pontuais gerados por movimento browniano fracionário. Utilizamos esses resultados para demonstrar que a entropia é consistente e sensível a variações nas estruturas espaciais e temporais, tanto em dados regulares quanto irregulares. Também descrevemos uma aplicação para análise de autenticidade de assinaturas, na qual usamos a distribuição de padrões ordinais (usada para calcular a entropia) para distinguir entre assinaturas genuínas e forjadas. No Capítulo 3, mostramos uma extensão do método que possibilita sua aplicação em imagens, testando-o em imagens de superfícies fractais e em imagens de cristais líquidos colestéricos. Por fim, no Capítulo 4, mostramos outra extensão do método que possibilita seu uso em séries temporais irregularmente amostradas.

A maioria dos testes e experimentos computacionais que realizamos são baseados em previsões por aprendizado de máquina, ou seja, treinamos um modelo usando uma parte dos dados e testamos se esse modelo consegue prever as propriedades usando a outra parte. Por isso, no Apêndice A, descrevemos os fundamentos básicos dos métodos de aprendizado de máquina que utilizamos.

CAPÍTULO 1

Entropia de permutação dos k-primeiros vizinhos

A entropia de permutação é uma medida adequada para distinção entre regimes aleatórios e regulares em séries temporais [2]. A ideia central dela é dividir a série em partições e, por meio de uma representação simbólica, identificar os padrões de ordenação nas regiões delimitadas pela partição. A partir da frequência relativa de ocorrência de cada padrão, cria-se uma distribuição de probabilidade de padrões ordinais, da qual se calcula a entropia de Shannon [33], obtendo assim a entropia de permutação. A aplicabilidade do método é notável por sua robustez e alta eficiência computacional.

A medida que iremos introduzir com nosso trabalho [32], a entropia de permutação dos k-primeiros vizinhos (*k-nearest neighbor permutation entropy*) ou entropia de permutação k-nn por brevidade, amplia os fundamentos da entropia de permutação para lidar com dados desestruturados, mantendo ao mesmo tempo suas características de robustez e eficiência computacional. Como veremos, o método envolve três etapas essenciais: *i)* criação de uma representação gráfica dos dados distribuídos espacialmente, transformando os pontos em nós e conectando os k-primeiros vizinhos de cada ponto; *ii)* amostragem de múltiplas séries temporais correspondentes à trajetória de caminhadas aleatórias no grafo, nas quais os valores da série corresponder aos valores associados aos nós visitados pela trajetória; *iii)* cálculo da entropia usando as séries temporais amostradas e a representação simbólica de Bandt-Pompe. Após essas etapas, obtém-se uma entropia S que, após normalizada, varia entre 0 e 1, sendo 0 indicativo de completa regularidade e 1 indicativo de total irregularidade na distribuição ordinal.

Neste capítulo, vamos detalhar essas três etapas e as ideias subjacentes a cada uma delas. Também iremos apresentar as limitações dos parâmetros utilizados, como ajustar esses parâmetros, o tipo de amostragem, a eficiência computacional e a representação gráfica

dos dados.

1.1 Representação gráfica e vizinhança

A nossa medida foi projetada para caracterizar dados que não possuem estrutura regular, isto é, um conjunto de pontos que estão dispersos de forma irregular no espaço e possuem valores associados a cada um deles. No caso mais geral, esses pontos pertencem a um espaço multidimensional. Podemos considerar que os pontos são caracterizados por um vetor de posição $\vec{r}_i$ e um valor associado ou estado z_i, para $i = 1, 2, \ldots, N$, com N sendo o tamanho do conjunto de dados. Por exemplo, vetores de posição bidimensionais $\vec{r}_i = (x_i, y_i)$ podem denotar as localizações espaciais do epicentro de um terremoto, enquanto os valores z_i podem representar magnitude do terremoto. Outro exemplo viável seria se a posição dos pontos representasse as localidades de cidades, enquanto os valores associados a cada ponto representassem algum indicador urbano, como a população ou o produto interno bruto.

O primeiro passo em nosso método envolve a construção de um grafo dos k-vizinhos mais próximos de cada ponto. Um grafo é um par (V, A) em que V é um conjunto e A é um subconjunto de pares de V, sendo os elementos de V chamados de vértices e os elementos de $A \subset V \times V$ chamados de arestas. Uma aresta é um par $\{i, j\}$ que liga dois vértices e, caso ligados, os vértices são ditos serem vizinhos ou adjacentes. Na linguagem de sistemas complexos é comum chamar o vértice de nó e o grafo de rede. Ao longo do nosso texto as duas nomenclaturas são usadas de maneira intercambiável. As arestas de uma rede podem ser representadas por uma matriz cujos os elementos A_{ij} indicam a presença ou ausência de uma aresta, isto é, $A_{ij} = 1$ se existir conexão entre i e j ou $A_{ij} = 0$ caso não existir. Essa representação da rede em forma matricial é chamada de matriz de adjacência da rede. Em uma rede direcionada A_{ij} indica uma aresta que tem origem em i e destino em j. Uma rede não direcionada pode ser pensada como possuindo uma aresta que corresponde a duas arestas direcionadas $\{i, j\} = \{(i, j), (j, i)\}$, um resultado direto disso é que a matriz de adjacência que representa redes não direcionadas é necessariamente simétrica ($A_{ij} = A_{ji}$). Elementos não nulos na diagonal são chamados de autoarestas (*autoloops*) e representam conexões que têm o mesmo nó como ponto de partida e de chegada.

Em nosso grafo criado para representar os dados (primeiro passo do método), os nós correspondem aos pontos individuais do dado e as arestas não direcionadas conectam os nós aos seus k vizinhos mais próximos com base nos vetores de posição $\vec{r}_i$. Embora qualquer métrica de distância seja aplicável, toda a nossa análise utiliza a distância euclidiana. A Figura 1.1 representa um conjunto de dados hipotético de $N = 19$ pontos dispostos irregularmente e ilustra os grafos de vizinhos mais próximos resultantes ao considerar $k = 2$ e $k = 3$.

O parâmetro k, que representa o número de vizinhos mais próximos, determina a estrutura do grafo e influencia se o foco será nas estruturas de dados locais ou globais. O exemplo

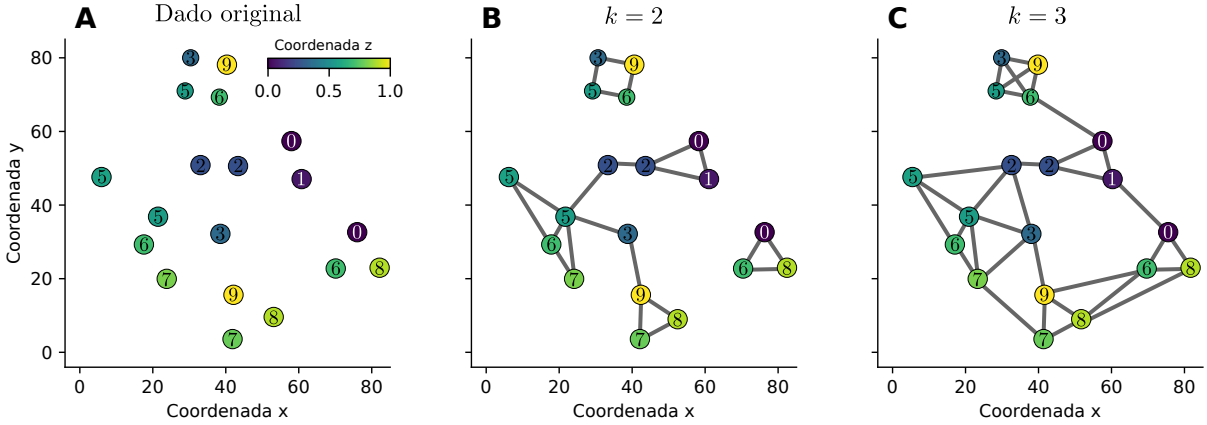

Figura 1.1: Grafo dos k-vizinhos mais próximos. O primeiro passo consiste em criar um grafo a partir de dados espalhados, sendo que cada ponto de dados z_i representa um nó, com arestas não direcionadas conectando pares $i \leftrightarrow j$ quando j está entre os k vizinhos mais próximos de i. (A) Ilustração de um conjunto de dados com pontos distribuídos irregularmente $\{z_i\}_{i=1,\ldots,N}$ no plano xy, no qual cada par de coordenadas (x_i, y_i) está associado a um valor z_i. A cor usada corresponde ao valor associado a cada ponto. (B) Representação a partir do grafo usando $k = 2$, no qual é possível observar que apenas os pontos mais próximos entre si foram conectados e que regiões muito separadas não possuem arestas que as conectam. (C) Representação a partir do grafo usando $k = 3$, nele pode-se observar que o grafo conecta todas as regiões e cria mais conexões nas regiões que já eram próximas.

da Figura 1.1 ilustra bem essa diferença. Quando $k = 2$, o grafo resultante consiste em várias componentes desconexas, formando três sub-grafos distintos no total. Um caminhante jamais poderá transitar de uma componente para outra, limitando-se a coletar informações apenas na ordem permitida pelas arestas de cada subgrafo. Nesse sentido, teremos uma amostragem mais local. Já quando $k = 3$, observa-se uma diferença significativa: o grafo resultante não possui partes desconexas, permitindo que um caminhante, dependendo do comprimento do percurso, colete informações mais abrangentes de diferentes regiões dos dados. Isso possibilita a percepção das diferenças entre as regiões do dado. Além disso, com mais conexões estabelecidas, as regiões que já eram próximas ficam ainda mais interconectadas, criando uma rede de caminhos dentro de cada região. Isso cria "atalhos" dentro das partes que já eram próximas, permitindo que um caminhante amostre informações diferentes. Os padrões obtidos nesse grafo mais conectado incluirão todos os padrões que seriam extraídos no caso de um grafo construído com menos vizinhos, já que o grafo criado com $k - 1$ estará sempre contido no grafo criado com k. No entanto, a proporção que cada padrão contribui para a distribuição de probabilidade pode mudar.

Existem limitações na escolha de k. Um valor excessivamente pequeno de k resulta em um grafo de vizinhos mais próximos com inúmeras componentes conectadas pequenas, confinando as trajetórias amostradas dentro desses componentes e, assim, limitando a iden-

tificação de padrões entre pontos distantes. Por outro lado, um valor excessivamente grande de k resulta em um grafo que se assemelha a uma estrutura totalmente conectada, tornando as trajetórias amostradas comparáveis a uma seleção aleatória de pontos.

Essa interação entre estrutura local e global é similar aos princípios fundamentais do método de aproximação e projeção uniforme de variedades (UMAP) [34,35], uma técnica de ponta para redução de dimensionalidade. O UMAP constrói uma versão ponderada de um grafo dos k-vizinhos mais próximos a partir de dados de alta dimensionalidade (chamado de complexo simplicial difuso [34]) e o projeta em um espaço de menor dimensionalidade usando um algoritmo de *layout* de grafo baseado em forças. Semelhante ao nosso método, a escolha de k no UMAP essencialmente determina até que ponto as estruturas de dados locais e globais são preservadas na projeção de baixa dimensionalidade. Esse processo de construção da rede poderia ser realizado exatamente da mesma forma que o método do UMAP, preservando assim a estrutura topológica dos dados. A rede resultante seria uma rede ponderada, isto é, cada ligação possuiria um peso indicando o quão fortemente dois nós estão relacionados entre si. A matriz de adjacência de uma rede ponderada possui elementos os A_{ij} nulos se uma aresta não existir entre os nós e, se existir uma aresta entre esses nós, o elemento de matriz pode assumir um valor real que corresponde ao peso da ligação, isto é, $A_{ij} \in \mathbb{R} \setminus \{0\}$. Para levar em conta esse peso associado às conexões no cálculo da entropia, poderíamos ponderar também a probabilidade de um caminhante aleatório na rede de acordo com esses peso, ou seja, uma conexão com peso maior seria mais provável de ser escolhida como caminho do que uma com peso menor. No entanto, para o propósito de calcular a entropia de permutação k-nn, nos restringimos apenas à criação do grafo a partir dos k-primeiros vizinhos, pois isso, como veremos, já é suficiente para capturar adequadamente a estrutura de diversos tipos de dados.

A representação gráfica a partir dos k-vizinhos mais próximos apresenta outra vantagem significativa: ela é independente da escala absoluta dos dados. Isso significa que, ao aplicá-la em conjuntos de dados com escalas completamente diferentes, a técnica trata ambos da mesma maneira. Ela cria um grafo que leva em consideração apenas a distância relativa entre cada ponto e seus vizinhos, permitindo que o método seja usado independentemente da escala dos dados. A única consideração relevante é o número de vizinhos, um parâmetro "universal" que influencia no tipo de representação obtida a partir dos dados.

Em relação a implementação computacional desse primeiro passo, existem algumas sutilezas. É comum usar a matriz de adjacência A_{ij} para representar um grafo. No entanto, optamos por uma representação do grafo que é mais eficiente computacionalmente, a *compressed sparse row* (CSR). Nela, um grafo com n nós e m conexões direcionadas é representado por uma lista de nós e uma lista de conexões. A lista que representa as conexões C, tem um tamanho m e é formada pela concatenação das listas de adjacência de todos os nós. Cada lista de adjacência de um nó é composta pelos índices dos nós aos quais ele está conectado.

Além disso, o índice de cada conexão direcionada (entre 0 e $m-1$) é mapeado ao índice do seu nó de chegada. Dessa forma, todas as conexões de um grafo são armazenadas continuamente. Essa característica melhora a localidade do armazenamento na memória, pois as conexões de saída de um vértice são armazenadas juntas e facilita sua identificação rápida. Em comparação, na representação matricial, deve-se fazer um algoritmo que verifique todos os elementos de uma dada linha e coluna da matriz para saber quais são as conexões de um nó. Lembrando que, no nosso caso, usamos grafos não direcionados e cada conexão é armazenada duas vezes na forma de uma conexão direcionada indo e outra voltando. A lista que representa os nós V possui um tamanho $n+1$, sendo que os n primeiros elementos estão na ordem do índice dos nós e o último elemento é o número total de conexões no grafo m. Essa lista é uma indexadora da lista C, pois cada índice de nó é mapeado no índice da primeira conexão que parte dele na lista de conexões. Os índices dos nós vão de 0 a $n-1$. Os elementos da lista são escolhidos de maneira que o número de conexões do i-ésimo nó seja igual a $V[i+1]-V[0]$.

Essa representação permite que conhecendo o índice do nó, tenha-se também o índice da primeira conexão e o número de elementos após o índice da conexão que são pertencentes àquele dado nó. Logo, obtêm-se todas as conexões que partem dele. Essa lógica é imprescindível para a realização de caminhadas aleatórias dentro de grafos, pois permite a rápida localização dos nós adjacentes em cada passo da caminhada. Tal fato tornou nosso código para calcular a entropia de permutação k-nn altamente otimizado.

1.2 Amostragem de trajetórias

De posse do grafo representativo dos dados, o próximo passo envolve a extração de séries temporais dos valores z_i. Para isso, utilizamos caminhadas aleatórias dentro do grafo. Amostramos os valores z_i iniciando n caminhadas aleatórias de comprimento w a partir de cada nó no grafo. Esses caminhantes aleatórios são intencionalmente direcionados para produzir uma amostragem em profundidade. Por amostragem em profundidade, nos referimos a uma caminhada que visite nós cada vez mais distantes do nó de origem.

Uma caminhada aleatória convencional dentro de uma rede tem uma probabilidade uniforme de se mover para qualquer vizinho do nó atual. Embora essa abordagem possa ser adequada para certos contextos, optar por uma caminhada que evita revisitar nós já percorridos oferece uma vantagem significativa, especialmente quando o objetivo é extrair padrões dos dados. Dentre a vasta gama de padrões possíveis, aqueles que revisam o mesmo nó várias vezes tendem a ser menos significativos. Em contrapartida, os padrões que cobrem uma extensa seção da rede são muito mais representativos e informativos sobre a estrutura dos dados.

Para obter essa amostragem intencionalmente direcionada usamos uma estratégia seme-

lhante ao `node2vec` [36], um algoritmo projetado para produzir representações vetoriais de nós de uma rede. O processo é uma caminhada aleatória de segunda ordem, na qual o caminhante decide se mover de sua posição atual b para uma posição subsequente c, considerando sua posição anterior a, de acordo com a probabilidade de transição não normalizada

$$\rho_{bc} = \begin{cases} 1/\lambda & \text{se } s_{ac} = 0 \\ 1 & \text{se } s_{ac} = 1 \\ 1/\beta & \text{se } s_{ac} = 2 \end{cases}, \tag{1.1}$$

na qual s_{ac} denota a distância do caminho mais curto entre os nós a (posição anterior) e c (posição subsequente), enquanto λ e β são parâmetros positivos que controlam o viés do caminhante.

De maneira prática, a primeira condição, $s_{ac} = 0$, só é possível se a posição subsequente for igual à posição anterior, ou seja, $a = c$. Isso representa a possibilidade do caminhante retornar ao nó anterior, dando um passo de b para a. Assim, um valor alto de λ, superior a 1, reduz a probabilidade do caminhante retornar à sua posição anterior a, enquanto valores mais baixos favorecem movimentos de retrocesso. A segunda condição, $s_{ac} = 1$, corresponde à possibilidade de dar um passo para um nó que seja vizinho de a, ou seja, entre os vizinhos de a. A terceira condição, $s_{ac} = 2$, corresponde a dar um passo para um nó que não é a nem um vizinho de a. Esses nós subsequentes são os vizinhos dos vizinhos de a, cuja menor distância é devida ao caminho composto pelas arestas $\{c, b\}$ e $\{b, a\}$. Assim, um valor alto de β, maior que 1, aumenta a preferência do caminhante se mover para nós próximos à sua posição anterior, enquanto valores mais baixos tornam o caminhante mais propenso a se mover para além da vizinhança imediata de sua localização anterior.

É essa dependência dos vizinhos do nó anterior que caracteriza a caminhada do `node2vec` como sendo de segunda ordem. Cada passo na sequência da caminhada é determinado não apenas pelo nó anterior b, mas também pelo nó anterior ao anterior a. Assim, ao escolher o próximo elemento na sequência da caminhada, representado pelo nó c, ele depende diretamente dos dois passos anteriores. Portanto, essa caminhada mantém uma memória de dois passos. Também é importante notar que, apesar da existência de viés, a natureza probabilística do processo impede que os caminhantes fiquem presos em quaisquer pontas soltas dentro do grafo.

A implementação deste procedimento em nosso código envolve dois sorteios. Primeiramente, é feita uma seleção entre os vizinhos do nó b. Em seguida, um segundo sorteio é realizado com base nos parâmetros λ e β, dependendo do caso do nó escolhido. No entanto, a ordem dos sorteios poderia ser invertida. Seria possível realizar primeiro um sorteio com base nos parâmetros para decidir qual conjunto de nós dentre os possíveis passos subsequentes escolher e, em seguida, fazer um segundo sorteio para determinar um nó dentro desse

conjunto. Isso ocorre porque a probabilidade de escolha do vizinho e a probabilidade de decidir se ele será o próximo, conforme definido na Equação 1.1, são independentes. Portanto, a probabilidade total do processo é o produto dessas probabilidades e elas podem ser realizadas em qualquer ordem devido à comutatividade. No entanto, esses métodos não são equivalentes em termos de eficiência computacional. O segundo método, que envolve realizar o sorteio dentro de cada conjunto, é consideravelmente mais custoso computacionalmente. Portanto, nosso código opta pelo primeiro método.

A escolha dos parâmetros λ e β possui uma significância mais profunda. Como já mencionamos, o método `node2vec` é uma técnica para representação vetorial de redes que utiliza caminhadas aleatórias como base para criar essas representações. O método parte da hipótese de homofilia [37], que postula que nós com alta interconexão e que pertencem a comunidades ou *clusters* semelhantes devem estar próximos uns dos outros na representação vetorial. Além disso, nós que desempenham funções estruturais semelhantes na rede também devem ser agrupados próximos na representação vetorial, conforme sugerido pela hipótese da equivalência estrutural [38]. A equivalência estrutural não está centrada na conectividade, o que significa que nós distantes na rede podem desempenhar funções estruturais similares. É importante ressaltar que essas duas características dos nós podem coexistir dentro de uma rede, resultando em uma complexa interação entre a estrutura da rede e as funções dos nós. Com base nessas características, o `node2vec` é um método que utiliza caminhadas aleatórias para permitir uma amostragem flexível, utilizando duas estratégias de busca. A primeira estratégia é conhecida como "amostragem em largura" (*breadth-first sampling*), na qual o conjunto de vizinhos é limitado aos nós imediatamente adjacentes à fonte, proporcionando uma visão detalhada da vizinhança de cada nó e da equivalência estrutural. A segunda estratégia é chamada de "amostragem em profundidade" (*depth-first sampling*), na qual a vizinhança é composta por nós amostrados sequencialmente a distâncias crescentes do nó de origem, refletindo uma visão mais ampla da vizinhança, o que é crucial para inferir comunidades com base na homofilia. Essas estratégias são determinadas pela escolha dos parâmetros λ e β na Equação 1.1.

Em todas as nossas aplicações, a menos que seja especificado diferente, utilizamos $\lambda = 10$ e $\beta = 0.001$ para garantir uma amostragem em profundidade dos valores z_i. A Figura 1.2 ilustra os dois tipos de amostragem e como cada um gera caminhadas distintas e, consequentemente, extrai padrões diferentes. Mais especificamente, o parâmetro β é o que controla se a caminhada vai ser mais sensível a homofilia ou a equivalência estrutural. Se $\beta > 1$, a caminhada aleatória tende a ser direcionada para nós próximos ao nó a no caminho, aproximando-se do comportamento de uma busca em largura no sentido de que nossas amostras consistem em nós dentro de uma pequena localidade, ou seja, uma visão mais local. Se $\beta < 1$, a caminhada tem maior probabilidade de visitar nós que estão mais distantes do nó a. Sendo assim, esse comportamento reflete uma busca em profundidade que enco-

raja a exploração para fora da vizinhança do nó a e proporciona uma visão mais global da rede. O parâmetro λ controla apenas a probabilidade de revisitar um nó que já faz parte da trajetória.

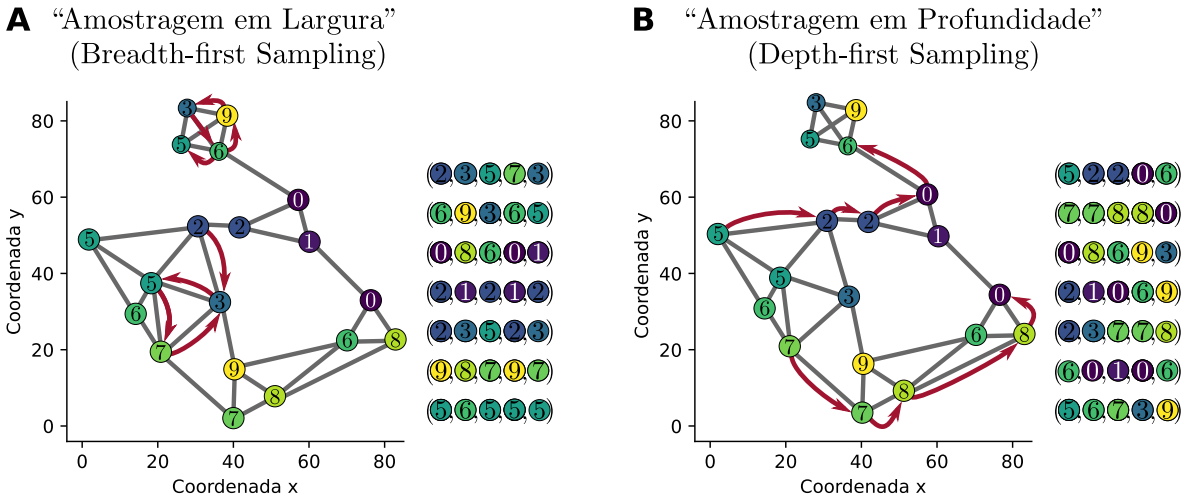

Figura 1.2: Amostragem enviesada. Dois tipos de amostragem usando diferentes parâmetros λ e β. A imagem mostra duas redes idênticas criadas a partir da conexão dos $k = 3$ primeiros vizinhos. (A) A amostragem em largura tende a visitar nós mais próximos dos seus vizinhos. Sendo assim, a caminhada tende a amostrar padrões mais locais. (B) A amostragem em profundidade tende a visitar nós cada vez mais distantes dos seus vizinhos. Nesse caso, os padrões amostrados serão significativamente diferentes pois tenderão a capturar a estrutura global do dado.

Também existe uma interação sutil entre os parâmetros do `node2vec` e o número de vizinhos utilizados para construir a rede. Ao lidar com um determinado valor de k, o uso de parâmetros que tendem a afastar a caminhada do seu vizinho permite a obtenção de padrões na escala limitada pelo k usado. O valor de k restringe o quão longe uma caminhada pode avançar, sendo assim, ao aumentar k aumenta-se também a possibilidade da caminhada explorar cada vez mais profundamente a rede. Isso ocorre porque os vizinhos mais próximos do nó anterior aumentarão com k e surgirão atalhos entre pontos distantes. Um número baixo de vizinhos resultará em amostragem em largura, independentemente da escolha dos parâmetros λ e β, pois as opções de passos serão limitadas. Por outro lado, um valor alto de k aumenta as possibilidades de escolha e, consequentemente, o viés da caminhada se torna mais evidente. Essa interação entre o número de vizinhos k e os parâmetros da caminhada λ e β introduz uma nova capacidade para a medida de entropia de permutação k-nn, que pode ser utilizada para obter padrões em diferentes escalas espaciais.

1.3 Cálculo da entropia

Após realizar as caminhadas descritas no segundo passo, temos um total de $N \times n$ trajetórias de comprimento w, com N e n denotando o tamanho do conjunto de dados e o número de caminhadas por nó, respectivamente. Essas trajetórias exploram as relações entre pontos de dados adjacentes e oferecem uma maneira natural de definir padrões ordinais entre os pontos. Tendo essas trajetórias ou séries temporais, podemos agora aplicar a metodologia da representação simbólica da entropia de permutação [2]. Esse procedimento de calcular a entropia de permutação das caminhadas é o terceiro passo para obtenção da entropia de permutação dos k-primeiros vizinhos. De maneira resumida, essa medida é a entropia de permutação usual calculada para séries temporais proveniente de caminhadas em um grafo, sendo que este grafo representa dados irregulares espalhados no espaço.

Para descrever melhor esse procedimento, representamos uma dada caminhada como $\{\tilde{z}_t\}_{t=1,\dots,w}$ e a segmentamos em partições sobrepostas

$$u_q = (\tilde{z}_q, \tilde{z}_{q+1}, \dots, \tilde{z}_{q+d-1}) \tag{1.2}$$

com d sendo o tamanho da partição ou a dimensão de *embedding* e $q = 1, \dots, w - d + 1$ os índices de cada partição. Seguindo o procedimento de Bandt e Pompe [2], para cada partição, determinamos a permutação $\pi_q = (r_0, r_1, \dots, r_{d-1})$ dos números de índice $(0, 1, \dots, d-1)$ que organizam as observações de u_q em ordem crescente. Cada uma das partições é definida pela desigualdade $\tilde{z}_{q+r_0} \leq \tilde{z}_{q+r_1} \leq \cdots \leq \tilde{z}_{q+r_{d-1}}$ e, no caso de valores iguais, a ordem de ocorrência é mantida (isto é, se $\tilde{z}_{q+r_{s-1}} = \tilde{z}_{q+r_s}$ para algum $s \in (1, \dots, d-1)$, então $r_{s-1} < r_s$). Esse procedimento é replicado em todas as trajetórias amostradas, gerando $M = N \times n \times (w-d+1)$ permutações simbólicas ou padrões ordinais.

A Figura 1.3A ilustra o procedimento de obtenção das sequencias simbólicas. Temos $n = 3$ caminhadas aleatórias de tamanho $w = 6$ por nó e uma dimensão de *embedding* $d = 3$. Considerando como exemplo a primeira caminhada que começou no nó com valor $z_i = 0$, ($\{\tilde{z}_t\}_{t=1,\dots,6} = \{0, 1, 2, 2, 5, 6\}$), temos que a primeira partição é $u_1 = (0, 1, 2)$ e, ordenando esses elementos em ordem crescentes, $0 \leq 1 \leq 2$ ou $\tilde{z}_{1+0} \leq \tilde{z}_{1+1} \leq \tilde{z}_{1+2}$, encontramos $\pi_1 = (0, 1, 2)$ que corresponde a um padrão monotonicamente crescente. Tomando agora o exemplo da segunda partição da segunda caminhada ($\{\tilde{z}_t\}_{t=1,\dots,6} = \{7, 8, 9, 3, 5, 6\}$), $u_2 = (8, 9, 3)$, temos que a ordenação dos elementos em ordem crescente é $3 \leq 8 \leq 9$ ou $\tilde{z}_{2+2} \leq \tilde{z}_{2+0} \leq \tilde{z}_{2+1}$, logo o padrão ordinal é $\pi_2 = (2, 0, 1)$, correspondendo a um crescimento seguido de um decrescimento.

Após agrupar todos os M símbolos de permutação $\{\pi_s\}_{s=1,\dots,M}$ obtidos das trajetórias amostradas, calculamos a probabilidade $p_j(\Pi_j)$ de cada símbolo de permutação possível Π_j,

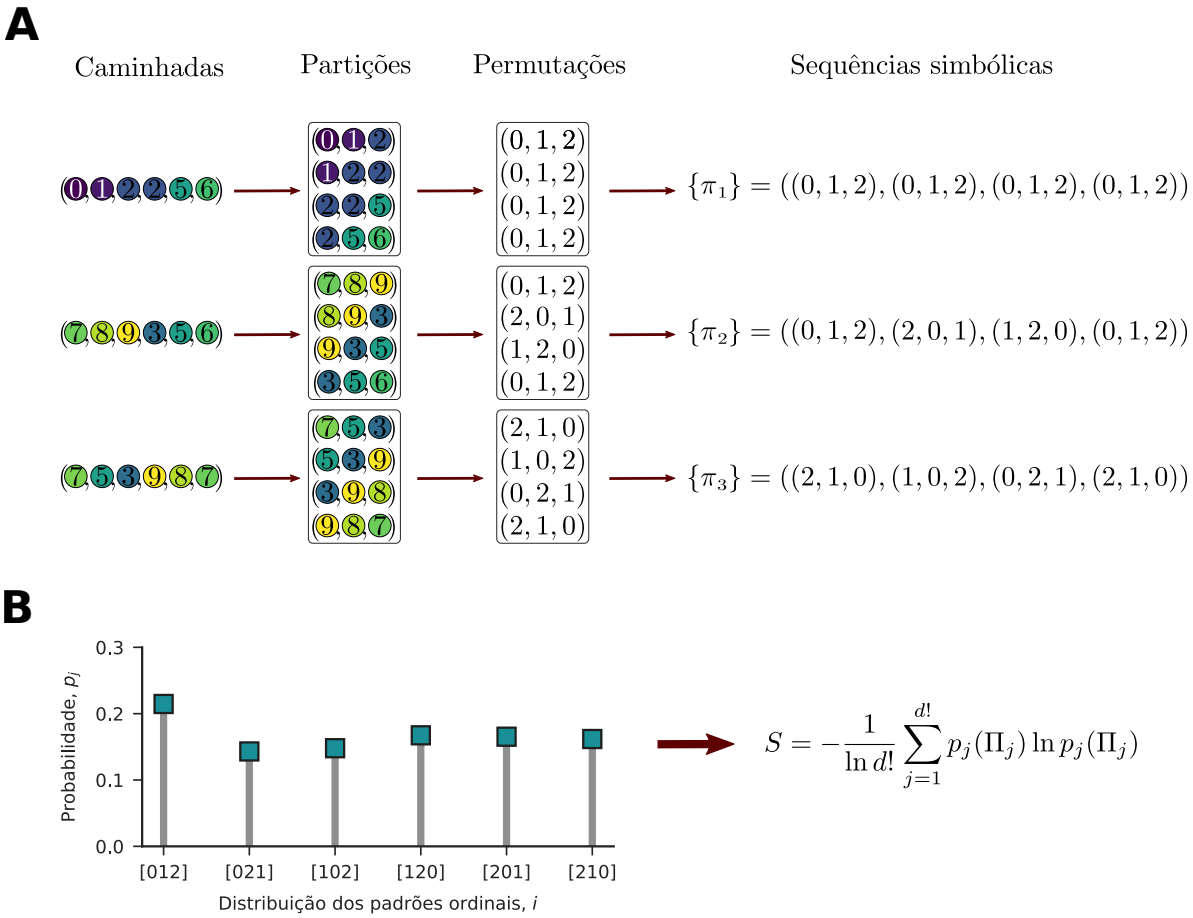

Figura 1.3: Procedimento de representação simbólica de Bandt-Pompe e cálculo da entropia de permutação k-nn. (A) A abordagem de Bandt-Pompe aplicada a três séries temporais de tamanho $w = 6$. O procedimento envolve a criação de partições sobrepostas de comprimento d (dimensão de *embedding*) e a organização dos índices das partições em ordem crescente de seus valores para determinar as permutações de ordenação de cada partição ($d = 3$ neste exemplo). (B) Distribuição ordinal correspondente a probabilidade de cada uma das $d!$ permutações possíveis e o cálculo de sua entropia de Shannon, que define assim a entropia de permutação dos k-primeiros vizinhos.

com $j = 1, \ldots, d!$, determinando sua frequência relativa

$$p_j(\Pi_j) = \frac{\text{total de permutações } \Pi_j \text{ em } \pi_s}{M}. \tag{1.3}$$

As probabilidades resultantes constituem a distribuição ordinal de probabilidades $P = \{\Pi_j\}_{j=1,\ldots,d!}$ e sua entropia normalizada de Shannon [33]

$$S = -\frac{1}{\ln d!}\sum_{j=1}^{d!} p_j(\Pi_j)\ln p_j(\Pi_j) \tag{1.4}$$

define a nossa entropia de permutação de k-vizinhos mais próximos (ou entropia de permu-

tação k-nn, por brevidade). Essa parte final do nosso método está ilustrada na Figura 1.3B para o caso das três caminhadas.

A entropia de permutação k-nn S mede a uniformidade da distribuição de probabilidade; quanto mais próximas forem as frequências de ocorrência de cada símbolo, mais próximo será o valor de S de seu máximo, isto é, $S \approx 1$. Em outras palavras, todos os $d!$ possíveis padrões ordinais são igualmente prováveis nos dados analisados. Por outro lado, quando $S \approx 0$, a distribuição dos símbolos é muito heterogênea, ou seja, existe um símbolo ou padrão ordinal que ocorre com frequência muito maior que a dos demais.

No caso de dados irregulares, os pontos de dados $\{\vec{r}_i, z_i\}$ podem gerar padrões pela combinação de dois processos diferentes. O primeiro é devido aos valores z_i e o segundo é devido à distribuição espacial dos pontos. Essas duas características influenciam no tipo de padrão que vai surgir, pois uma caminhada vai amostrar os valores z_i apenas na ordem que é limitada pelas possíveis trajetórias do grafo. Assim, por exemplo, não basta os valores z_i serem regulares; eles também precisam estar dispostos espacialmente de uma maneira regular para que a entropia seja baixa. Nesse sentido, uma outra maneira de interpretar os limites da entropia é imaginar que eles quantificam o grau de irregularidade na distribuição dos pontos de dados $\{\vec{r}_i, z_i\}$. Esperamos encontrar $S \approx 0$ (limite inferior) quando os valores adjacentes de z_i apresentarem uma configuração regular caracterizada pela predominância de um único padrão ordinal. Por outro lado, $S \approx 1$ (limite superior) sugere a ausência de estrutura regular entre os valores adjacentes de z_i, indicando a falta de preferência por padrões ordinais específicos.

Nossa abordagem incorpora ainda o número de caminhadas aleatórias por nó n e o comprimento dessas caminhadas w como parâmetros. Juntamente com o tamanho do conjunto de dados N e a dimensão de *embedding* d, esses dois parâmetros especificam o número total de símbolos de permutação $M = N \times n \times (w - d + 1)$ extraídos dos dados. Portanto, é essencial ter $M \gg d!$ para obter uma estimativa confiável da probabilidade de todos os $d!$ padrões ordinais possíveis. Naturalmente, w deve exceder d para acomodar as partições u_q dentro das trajetórias amostradas. No entanto, a escolha entre um grande número de caminhadas relativamente curtas por nó versus um pequeno número de caminhadas longas por nó tende a ter impacto mínimo na estimativa da entropia de permutação k-nn, uma vez que os caminhantes aleatórios lembram apenas de sua posição imediatamente anterior. Além disso, cada réplica das caminhadas aleatórias gera trajetórias amostradas e símbolos de permutação distintos, resultando em diferentes estimativas das distribuições ordinais e, consequentemente, em diferentes valores para a entropia de permutação k-nn S. Portanto, além de garantir que M exceda significativamente $d!$, as caminhadas aleatórias devem fornecer uma amostra representativa dos possíveis padrões ordinais extraídos do grafo, tornando as variações em S desprezíveis para a análise. O número total de caminhadas de comprimento d pode ser calculado usando potências da matriz de adjacência do grafo e pode servir como

um guia para definir os valores de n e w. Porém, uma estratégia alternativa mais eficaz é aumentar incrementalmente o número de caminhadas até alcançar um nível desejável de estabilidade na entropia de permutação k-nn.

CAPÍTULO 2

Caracterização de dados irregulares

A entropia de permutação k-nn foi concebida para lidar com dados irregulares como os geoespaciais, nos quais os pontos estão em duas dimensões e um valor é atribuído a cada ponto, isto é, um vetor de posição $\vec{r}_i$ e um valor associado z_i. Matematicamente, esse tipo de dado também é conhecido como processo pontual (*point process*) [39].

Distinguir entre padrões regulares e irregulares em dados espalhados em duas dimensões constitui um dos principais empreendimentos do trabalho presente. Sendo assim, nosso objetivo é demonstrar a capacidade da medida de entropia de permutação k-nn. Para isso, utilizamos dois grandes conjuntos de testes e realizamos uma análise extensiva neles. O primeiro conjunto corresponde a um movimento Browniano Fracionário cuja estrutura espacial é fixa, enquanto a estrutura dos valores z_i varia. No segundo conjunto, a estrutura dos valores é constante, enquanto a estrutura espacial varia. Nos dois casos, verificamos a capacidade da entropia de permutação k-nn para prever parâmetros-chave desses dados utilizando técnicas de aprendizado de máquina, além de compararmos nossos resultado com outras medidas disponíveis na literatura.

Além disso, com o objetivo de avaliar nossa medida de entropia em dados empíricos, analisamos trajetórias referentes ao deslocamento da ponta de uma caneta em um conjunto de assinaturas. Nesse caso, empregamos a distribuição dos padrões ordinais e, por meio de procedimentos de aprendizagem de máquina, classificamos as assinaturas em genuínas ou forjadas, comparando nossos resultados com uma técnica similar baseada na distribuição ordinal proveniente do procedimento usual de Bandt-Pompe [40].

2.1 Dados simulados com estrutura espacial fixa

Nosso primeiro conjunto de dados irregulares z_i espalhados em duas dimensões $\vec{r}_i = (x_i, y_i)$ é definido como

$$\begin{aligned} x_i &= x_{i-1} + \xi_{h_x} \\ y_i &= x_{i-1} + \xi_{h_y} \,, \\ z_i &= z_{i-1} + \xi_{h_z} \end{aligned} \tag{2.1}$$

na qual ξ_{h_x}, ξ_{h_y} e ξ_{h_z} representam ruídos Gaussianos fracionários [41, 42] com média zero, variância unitária e expoentes de Hurst h_x, h_y e h_z, respectivamente. Esses termos de ruído são simulados numericamente usando o método de Hosking [43]. Desse modo, as variáveis x_i, y_i e z_i denotam movimentos Brownianos fracionários, com x_i e y_i servindo como as coordenadas dos dados e z_i como o valor associado. Os expoentes de Hurst modulam a rugosidade desses processos estocásticos. Expoentes menores que $1/2$ resultam em séries temporais exibindo alternâncias mais frequentes nos sinais do incremento do que o esperado por acaso (comportamento anti-persistente). Além disso, a série nesse intervalo tem um aspecto mais irregular e se assemelha a uma trajetória mais rugosa. Por outro lado, expoentes maiores que $1/2$ levam a séries temporais em que os incrementos mantêm seus sinais mais frequentemente do que o esperado por acaso (comportamento persistente) e a série resultante se assemelha a uma trajetória mais suave e regular. Quando o expoente de Hurst é $1/2$, essas séries correspondem ao movimento Browniano convencional. Outra maneira de interpretar esses diferentes regimes devido ao expoente é por meio de sua dimensão fractal. A dimensão de cada série é igual a $2 - h_{x,y,z}$; portanto, no limite inferior, a trajetória parece ocupar o espaço de uma linha, enquanto que no limite superior, ela parece ocupar o espaço de um plano.

Escolhemos o movimento browniano fracionário pois ele é um modelo que já foi extensivamente estudado e possui diversas aplicações. Por exemplo, no campo da biologia celular, esse processo demonstrou ser um modelo para descrever o movimento sub-difusivo dos telômeros, estruturas que protegem as extremidades dos cromossomos, no núcleo de células humanas vivas [44, 45], e dos loci cromossômicos em células bacterianas vivas [46, 47]. Esse modelo também é utilizado para descrever o movimento de grânulos lipídicos em células mitóticas iniciais [48]. Assim, o movimento browniano fracionário é um modelo paradigmático que explica diversos fenômenos, de modo que qualquer medida com pretensão de ser utilizável em contextos gerais deve ser robusta para caracterizá-lo. Além disso, uma vez demonstrada essa utilidade no modelo, a medida tem o potencial de ser aplicável a todos os casos citados.

Consideramos inicialmente o caso em que os expoentes de Hurst para as coordenadas espaciais estão fixos em $h_x = h_y = 1/2$, enquanto o expoente h_z varia de 0.1 a 0.9 com incrementos de tamanho 0.1. Essas escolhas mantêm uma estrutura espacial constante na

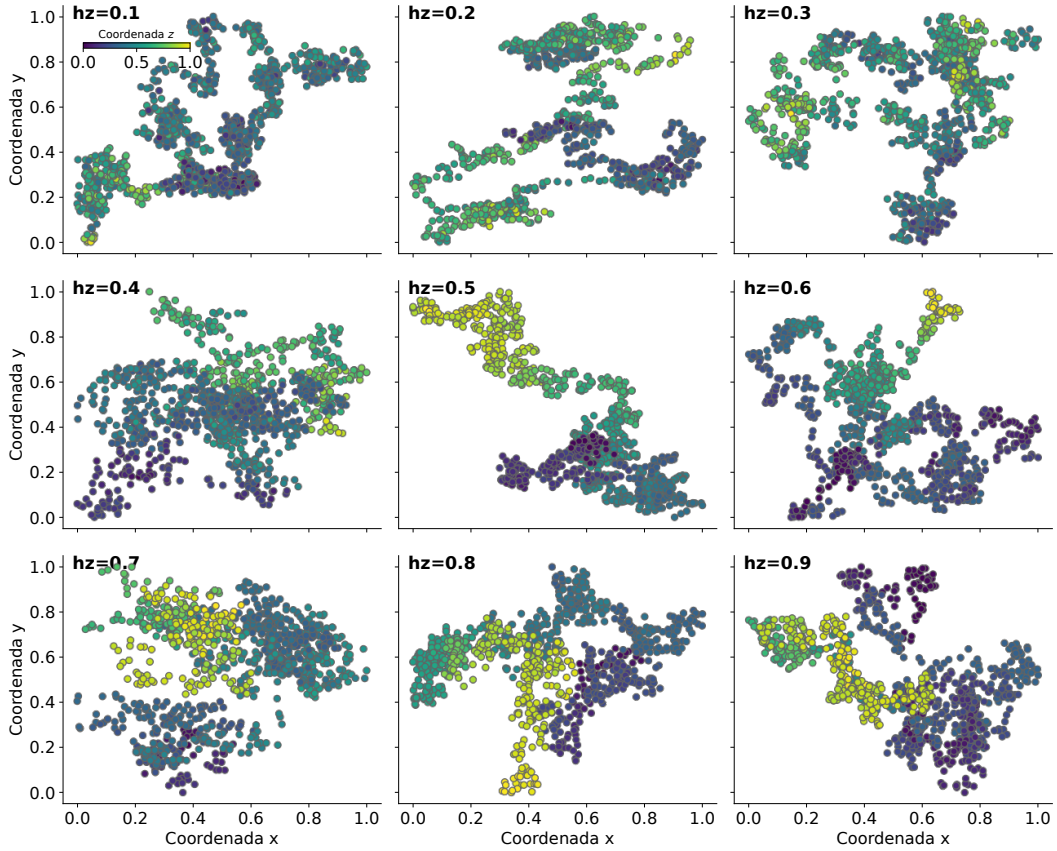

Figura 2.1: Movimento browniano fracionário com a estrutura da disposição espacial fixa. Ilustrações de conjuntos de dados distribuídos irregularmente nos quais as coordenadas x e y seguem um movimento browniano usual ($h_{xy} = 0.5$) e a coordenada z surge de um movimento browniano fracionário caracterizado por diferentes expoentes de Hurst h_z (conforme indicado nos gráficos).

distribuição de pontos e permite investigar se a entropia de permutação k-nn é capaz de identificar variações de regularidade nos valores z_i. A Figura 2.1 mostra nove exemplos de dados simulados para diferentes valores de h_z. Valores menores de h_z geram padrões mais aleatórios, enquanto valores mais altos produzem padrões espaciais caracterizados por valores semelhantes e tendências espaciais entre pontos adjacentes. Outra maneira de identificar esses regimes é observando as cores. No regime regular, pontos de mesma cor tendem a ocupar as mesmas regiões, formando um gradiente de cores bem definido ao longo da amostra. No regime irregular, as cores estão misturadas e nenhum gradiente pode ser observado.

Criamos um conjunto composto por cem réplicas independentes desses conjuntos de dados para cada h_z e diferentes tamanhos de conjunto de dados $N = 2^8, 2^9, \ldots, 2^{13}$. Usando esses dados, avaliamos como a entropia de permutação k-nn S depende do expoente de Hurst h_z para diferentes dimensões de *embedding* d, conforme ilustra o painel à esquerda da Figura 2.2A para $N = 1024$ e $d = 5$. O número de vizinhos usado para criar o grafo foi $k = 25$. Nesta figura, os marcadores correspondem aos valores médios de S e as áreas sombreadas

representam o intervalo de confiança de um desvio padrão. A entropia de permutação k-nn diminui monotonicamente com o expoente de Hurst h_z e, portanto, pode distinguir entre diferentes graus de estrutura espacial nos dados. Valores de entropia mais altos correspondem a uma maior aleatoriedade, enquanto valores mais baixos refletem padrões espaciais persistentes associados aos expoentes de Hurst mais altos.

Comparamos nossos resultados com o índice I de Moran [49], que é uma medida amplamente utilizada para quantificar autocorrelação espacial [50]. O I de Moran é essencialmente uma extensão ponderada espacialmente do coeficiente de correlação de Pearson que pode ser definida como

$$I = \frac{N}{W_0} \frac{\sum_i^N \sum_j^N \omega_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i^N (z_i - \bar{z})^2}, \tag{2.2}$$

na qual $\bar{z}$ denota a média de z_i, ω_{ij} é um peso entre os pontos de dados z_i e z_j, e $W_0 = \sum_i^N \sum_j^N \omega_{ij}$ é uma constante de normalização. O I de Moran geralmente varia de -1 a 1, com $I \to I_0 = -1/(N-1)$ para dados espaciais não correlacionados, enquanto $I > I_0$ e $I < I_0$ indicam, respectivamente, correlações espaciais positivas e negativas [51]. A forma mais simples para a matriz de pesos define $\omega_{ij} = 1$ para vizinhos imediatos e $\omega_{ij} = 0$ para outros pares de pontos (com $\omega_{ii} = 0$). Uma outra abordagem comum envolve definir o peso ω_{ij} por uma função de decaimento com a distância, por exemplo, $\omega_{ij} = 1/d_{ij}^{\alpha}$, com $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ sendo a distância entre os pontos z_i e z_j, e α um expoente que modula o decaimento [52]. Para nossa comparação, adotamos esta última abordagem e usamos diferentes valores de α para examinar como I depende do expoente de Hurst h_z. Conforme ilustrada o painel à direita da Figura 2.2A para $\alpha = 1$ e $N = 1024$, os valores de I tendem a aumentar com h_z. No entanto, diferentemente da entropia de permutação k-nn S, a relação exibida por I não é monótona e atinge um platô em torno de $h_z = 0.7$. Além disso, os valores de I apresentam uma dispersão relativa consideravelmente maior do que S, como evidenciado pela extensa área sombreada que representa o intervalo de confiança de um desvio padrão. Essas observações indicam que o I de Moran é menos eficaz na identificação dos diferentes graus de estrutura espacial de nossos dados simulados em comparação com a entropia de permutação k-nn S.

Também avaliamos a eficácia preditiva de S e I em tarefas de aprendizado de máquina empregando o classificador de vizinhos mais próximos [53] (mais detalhes do Apêndice A) para prever os valores de h_z usando cada medida como uma característica preditiva. Dividimos os dados em conjuntos de teste (20%) e treinamento estratificando pelos valores de h_z e treinamos os classificadores usando uma abordagem de validação cruzada com três partições (*three-fold cross-validation*) para optimizar o número de vizinhos no algoritmo de aprendizado. Em seguida, determinamos a acurácia das previsões no conjunto de teste. Repetimos a divisão entre treino e teste e o processo de treinamento em dez instâncias independentes, calculando a média e o desvio padrão da acurácia nos conjuntos de teste. A Figura 2.2B

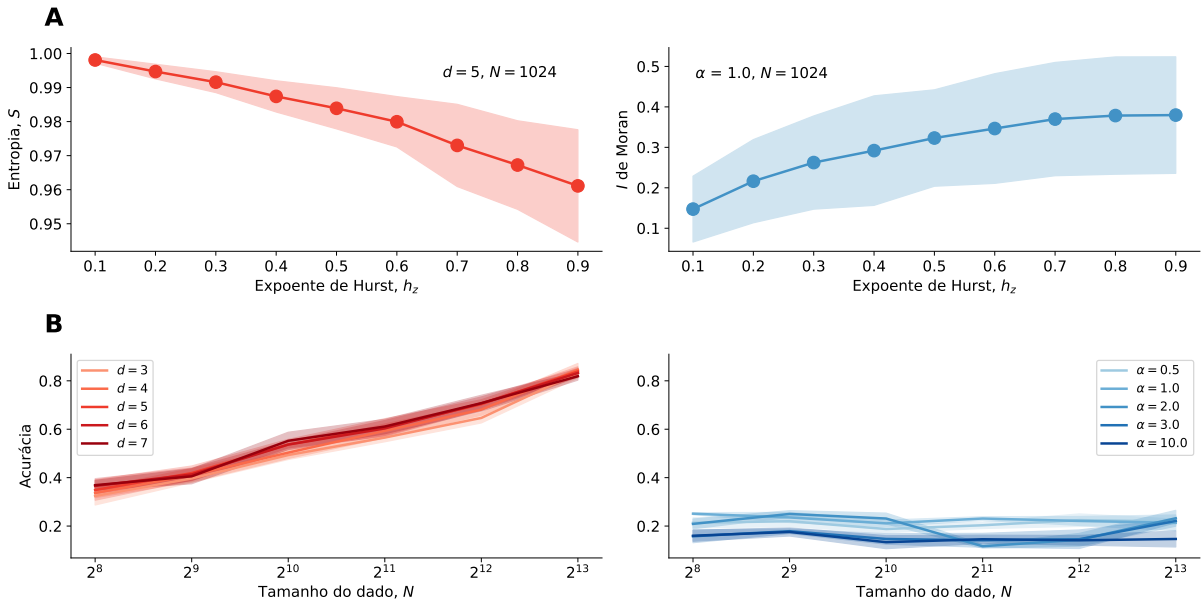

Figura 2.2: Comparação entre a entropia de permutação k-nn e o I de Moran na caracterização de movimento browniano fracionário. (A) Em vermelho temos a relação entre a entropia de permutação k-nn S (com $d = 5$) e o expoente de Hurst h_z. Em azul temos a dependência do índice I de Moran (com $\alpha = 1$) em relação ao expoente de Hurst h_z. Os marcadores representam os valores médios calculados a partir de cem réplicas independentes do processo que gera os conjuntos de dados (com $N = 1024$ pontos de dados) para cada $h_z \in \{0.1, 0.2, \ldots, 0.9\}$, enquanto as áreas sombreadas representam intervalos de confiança de um desvio padrão. (B) Acurácia das tarefas de classificação voltadas para a previsão de h_z usando a entropia de permutação k-nn S (tons de vermelho indicam diferentes dimensões de *embedding* d) e o I de Moran (tons de azul indicam diferentes valores de α) em função do tamanho do conjunto de dados N. As áreas sombreadas representam o desvio padrão dos níveis médios de acurácia estimados a partir de dez realizações independentes do processo de treinamento do classificador de vizinhos mais próximos.

mostra a acurácia média em função do tamanho do conjunto de dados obtida usando os valores de entropia para diferentes dimensões de *embedding* d (painel à esquerda), comparados a acurácia obtida a partir dos valores do I de Moran para vários valores de α. Esses resultados indicam que a entropia de permutação k-nn apresenta uma acurácia significativamente maior nas previsões do expoente de Hurst h_z em comparação com o I de Moran (que geralmente fica abaixo do limiar de acurácia de 20%).

Além disso, calculamos a matriz de confusão para essas tarefas de classificação nos conjuntos de teste, como mostrado na Figura 2.3A. A faixa diagonal na matriz indica que classificações incorretas feitas pelo algoritmo treinado com valores de S geralmente produzem expoentes de Hurst próximos ao valor real. Por outro lado, a matriz de confusão do algoritmo treinado com valores de I carece de um padrão diagonal semelhante, destacando a capacidade preditiva inferior de I em relação à entropia de permutação k-nn.

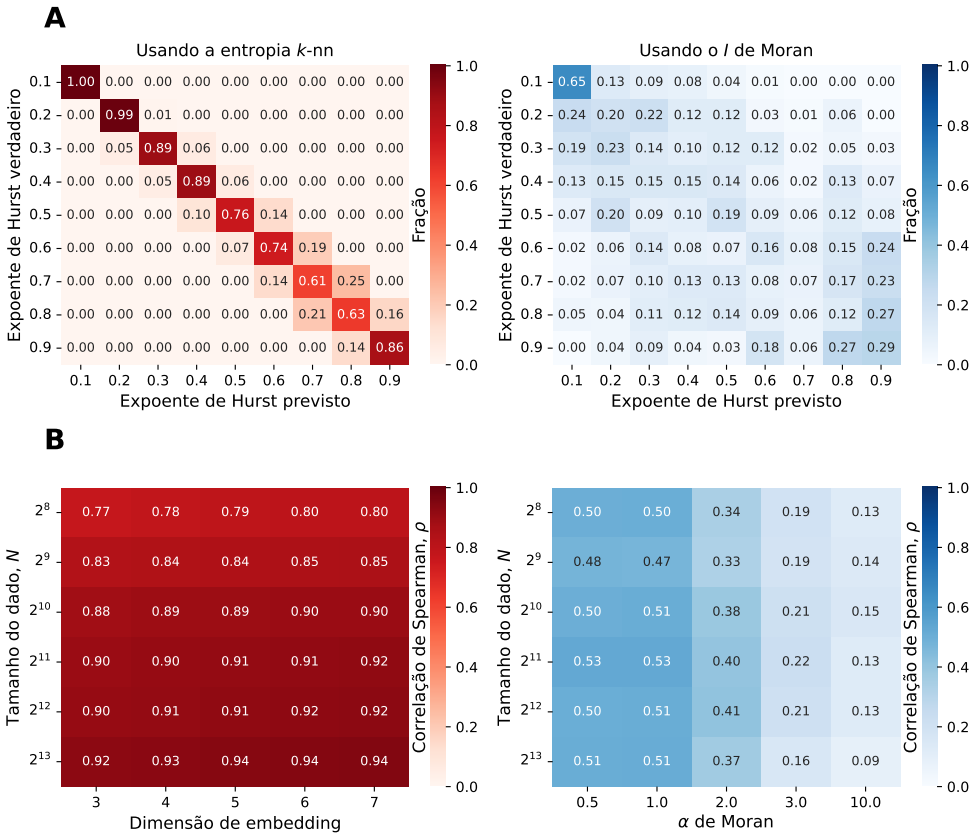

Figura 2.3: Detalhes das previsões e da acurácia ao empregar a entropia de permutação k-nn e o I de Moran. (A) Exemplos de matrizes de confusão resultantes da aplicação do algoritmo de aprendizado para prever o valor de h_z usando a entropia S com $d = 3$ (tons de vermelho) e as estimativas baseadas no I de Moran com $\alpha = 1$ (tons de azul). (B)Correlação de Spearman entre os valores de S e h_z (tons de vermelho) para diferentes tamanhos de conjuntos de dados (N, linhas) e dimensões de *embedding* (d, colunas), assim como a correlação entre os valores de I e h_z (tons de azul) para diferentes tamanhos de conjuntos de dados (N, linhas) e expoentes de distância (α, colunas).

Para comparar sistematicamente ambas as medidas e analisar se elas apresentam uma relação clara com o expoente de Hurst, usamos a medida de correlação de Spearman (ρ). Especificamente, calculamos a correlação entre S e h_z, organizando os resultados por tamanho do conjunto de dados N e dimensão de *embedding* d. Em seguida, comparamos esses achados com a correlação entre I e h_z, categorizada por tamanho do conjunto de dados N e expoente de distância α. Conforme mostra a Figura 2.3B, as correlações entre S e h_z são significativamente mais fortes do que aquelas entre I e h_z, independentemente do tamanho do conjunto de dados, dimensão de *embedding* ou expoente de distância.

Esses resultados corroboram nossa hipótese de que a entropia de permutação k-nn é sensível à estrutura dos valores z_i e consegue distinguir entre regimes regulares e irregulares.

2.2 Dados simulados com estrutura de valores fixa

Em uma segunda aplicação envolvendo dados não estruturados, investigamos se a entropia de permutação k-nn pode identificar mudanças estruturais na distribuição espacial dos pontos dos dados quando o padrão associado aos valores z_i permanece constante. Para isso, adaptamos o modelo apresentado na Equação 2.1 para criar conjuntos de dados espaciais $\vec{r}_i = (x_i, y_i)$ e z_i da seguinte maneira:

$$\begin{aligned} x_i &= x_{i-1} + \xi^{(x)}_{h_{xy}} \\ y_i &= y_{i-1} + \xi^{(y)}_{h_{xy}} \\ z_i &= i \end{aligned} , \tag{2.3}$$

na qual $\xi^{(x)}_{h_{xy}}$ e $\xi^{(y)}_{h_{xy}}$ são ruídos gaussianos fracionários [41, 42] com média zero, variância unitária e um mesmo expoente de Hurst h_{xy}.

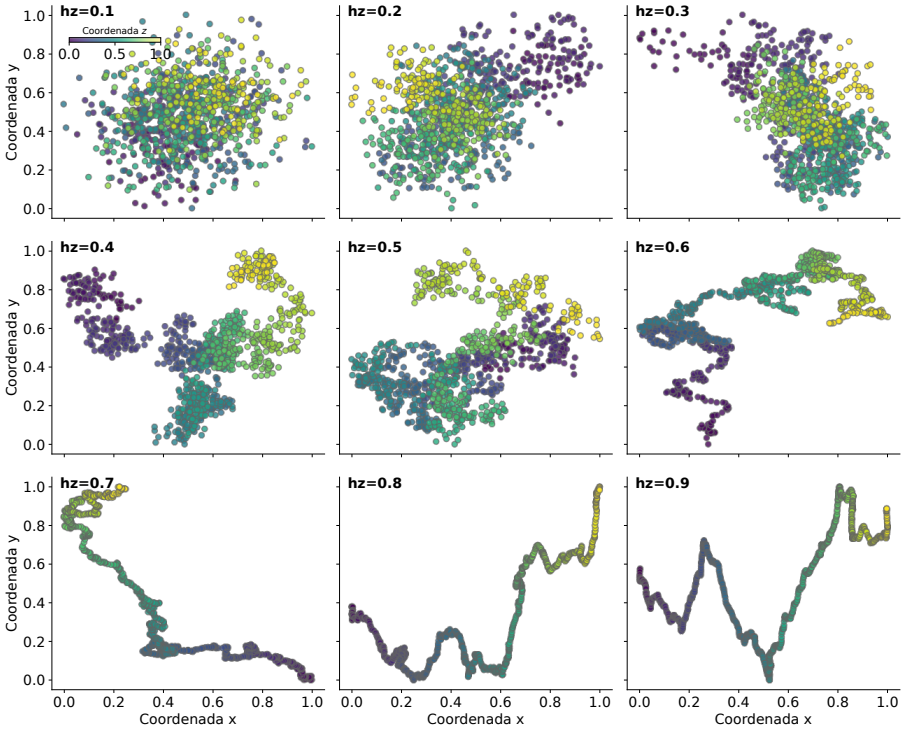

Figura 2.4: Movimento browniano fracionário com a estrutura de valores z_i fixa. Ilustração de dados espalhados gerados usando o movimento browniano fracionário descrito na Equação 2.3 com diferentes valores do expoente de Hurst h_{xy}.

Os conjuntos de dados gerados por meio deste modelo produzem um movimento browniano fracionário bidimensional nas coordenadas dos dados $\vec{r}_i = (x_i, y_i)$, enquanto os valores z_i simplesmente representam os índices dos passos temporais. Como ilustrado na Figura 2.4, valores menores de h_{xy} resultam em comportamento anti-persistente nas coordenadas dos dados, levando a padrões mais irregulares associados aos valores de z_i. Por outro lado, va-

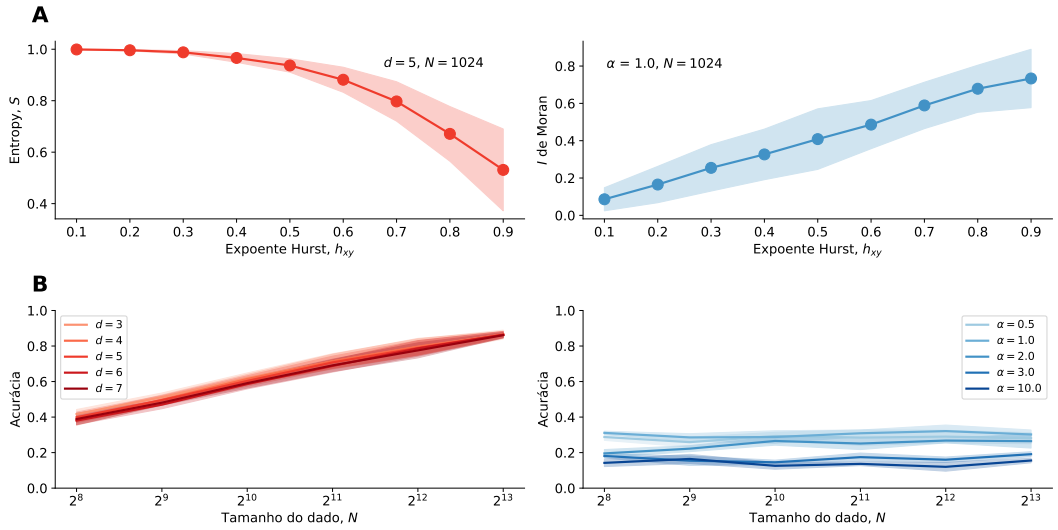

Figura 2.5: Comparação entre a entropia de permutação k-nn e o I de Moran. (A) Em vermelho temos a relação entre a entropia de permutação k-nn S (com $d = 5$) e o expoente de Hurst h_{xy}. Em azul temos a dependência do índice I de Moran(com $\alpha = 1$) em relação ao expoente de Hurst h_{xy}. Os marcadores representam os valores médios calculados a partir de cem réplicas independentes do processo que gera os conjuntos de dados (com $N = 1024$ pontos de dados) para cada $h_{xy} \in 0.1, 0.2, \ldots, 0.9$, enquanto as áreas sombreadas representam os intervalos de confiança de um desvio padrão. (B) Acurácia das tarefas de classificação voltadas para a previsão de h_{xy} usando a entropia de permutação k-nn S (tons de vermelho indicam diferentes dimensões de *embedding* d) e o I de Moran (tons de azul indicam diferentes valores de α), ambas em função do tamanho do conjunto de dados N. As áreas sombreadas representam um desvio padrão dos níveis médios de acurácia estimados a partir de dez realizações independentes do processo de treinamento do classificador de vizinhos mais próximos.

lores maiores de h_{xy} induzem um comportamento persistente nas coordenadas dos dados, resultando em padrões mais regulares caracterizados por valores adjacentes de z_i com tendências crescentes ou decrescentes. Do ponto de vista da dimensão fractal do movimento browniano, o regime irregular pode ser entendido como uma estrutura que se aproxima de um plano. Isso fica claro ao observar os pontos quando h_{xy} é baixo, pois eles estão espalhados de maneira a ocupar quase todo o espaço do plano. No regime regular, a dimensão fractal é baixa e processo se parece com uma linha, o que também pode ser visto quando h_{xy} é alto.

Consideramos $h_{xy} = 0.1, 0.2, \ldots, 0.9$ para gerar um conjunto com cem réplicas independentes desses processos para cada h_{xy}, além de variar o tamanho do conjunto de dados com $N = 2^8, 2^9, \ldots, 2^{13}$. Usando esses dados simulados, estimamos a entropia de permutação k-nn S em função do expoente de Hurst h_{xy} para várias dimensões de *embedding* d e analisamos como o I de Moran depende de h_{xy} para diferentes valores do expoente de distância α. A Figura 2.5A mostra essas relações para $N = 1024$ com valores de entropia calculados usando $d = 5$ e valores de I determinados com $\alpha = 1$. Consistente com nossas expectativas, notamos que S diminui com o aumento de h_{xy}, enquanto I aumenta. O I de Moran exibe uma associ-

ação aproximadamente linear, enquanto a entropia de permutação k-nn S exibe uma relação não linear marcada por uma rápida diminuição para $h_{xy} > 0.5$. Notavelmente, a dispersão relativa quantificada pelos intervalos de confiança de um desvio padrão é consideravelmente menor para S do que para I, especialmente para expoentes de Hurst menores.

Para comparar melhor as duas medidas, usamos a mesma estratégia da seção anterior e avaliamos a eficácia preditiva de S e I em tarefas de aprendizado de máquina usando o classificado de vizinhos mais próximos para prever o valor de h_{xy}. A Figura 2.5B exibe a acurácia média dessas tarefas preditivas em função do tamanho do conjunto de dados, usando valores de entropia S para diferentes dimensões de *embedding* e I de Moran calculados para diversos expoentes de distância. Mais uma vez, o desempenho da entropia de permutação k-nn supera significativamente o I de Moran, independentemente do tamanho do conjunto de dados ou dos parâmetros d e α.

Essa superioridade é ainda apoiada pelas matrizes de confusão da Figura 2.6A, nas quais observamos que previsões incorretas feitas pelo algoritmo treinado com valores de entropia estão tipicamente a um passo do verdadeiro expoente de Hurst, resultando em um padrão diagonal quase perfeito. Por outro lado, o algoritmo treinado com I de Moran apresenta uma faixa diagonal consideravelmente mais ampla, refletindo a menor eficácia dessa medida para a tarefa. Avaliamos também a qualidade das relações S versus h_{xy} e I versus h_{xy} calculando a correlação de Spearman e categorizando os resultados por N e também por d (para a entropia) e α (para o I de Moran). A Figura 2.6B apresenta essas correlações, indicando que a associação entre S e h_{xy} é substancialmente mais correlacionada do que a relação entre I e h_{xy}, independentemente do tamanho do conjunto de dados ou dos parâmetros d e α.

Nesta extensa análise, verificamos que a entropia também é sensível à estrutura espacial dos dados e consegue distinguir entre regimes regulares e irregulares. Somando com resultados obtidos na seção anterior, concluímos a eficácia geral da entropia. Devemos ressaltar que esses testes foram feitos deixando cada uma das estruturas fixas e variando a outra. Muito provavelmente um dado empírico não será tão bem comportado, e inferir se a irregularidade é proveniente da estrutura espacial ou dos valores z_i não é possível apenas olhando a entropia.

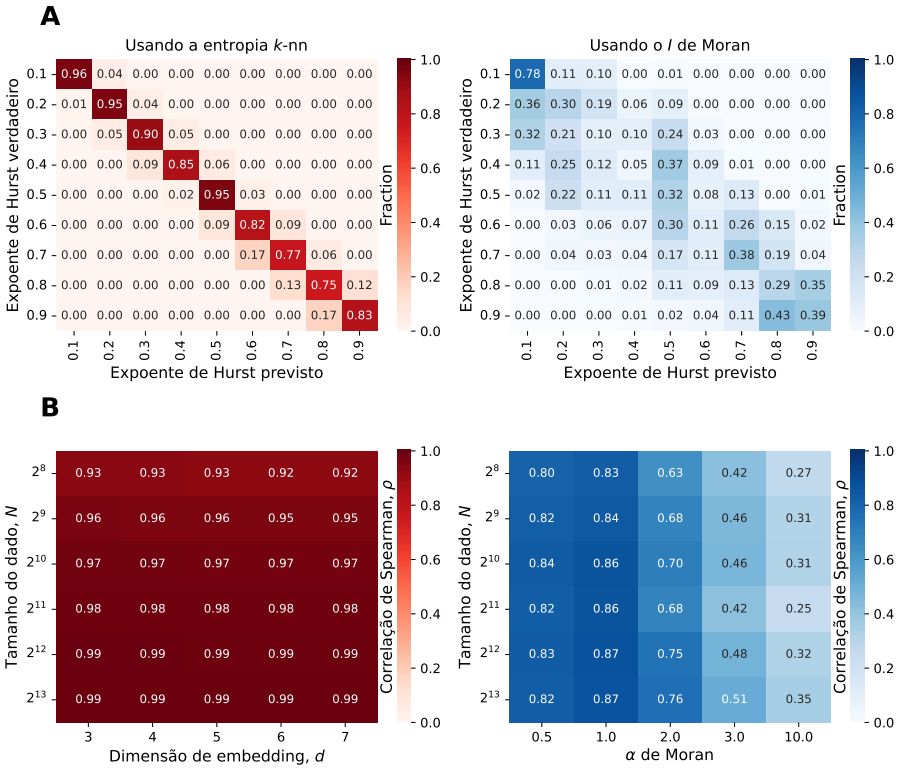

Figura 2.6: Detalhes das previsões e da acurácia das medidas de entropia de permutação k-nn e do I de Moran. (A) Exemplo de matrizes de confusão resultantes da aplicação do algoritmo de aprendizado usado para prever o valor de h_{xy} usando a entropia S para $d = 3$ (tons de vermelho) e as estimativas baseadas no I de Moran (tons de azul) (B) Correlação de Spearman entre os valores de S e h_{xy} (tons de vermelho) para diferentes tamanhos de conjuntos de dados (N, linhas) e dimensões de *embedding* (d, colunas), assim como a correlação entre os valores de I e h_{xy} (tons de azul) para diferentes tamanhos de conjuntos de dados (N, linhas) e expoentes de distância (α, colunas).

2.3 Classificação de assinaturas

Como uma primeira aplicação de nossa abordagem a dados empíricos, analisamos um conjunto de dados assinaturas referentes a "Primeira Competição Internacional de Verificação de Assinaturas (SVC2004)" [54]. Esse dado consiste em 40 conjuntos de assinaturas de pessoas, cada um correspondente a 20 assinaturas genuínas e 20 falsificações habilidosas, totalizando 1600 assinaturas. Os dados incluem assinaturas chinesas e ocidentais.

Nesta competição, cada participante produziu 20 assinaturas genuínas e 20 falsificações habilidosas de outras assinaturas em duas sessões. As assinaturas foram registradas em um mesa digitalizadora *Wacom Intuos*. Por questões de privacidade, os participantes criaram novas assinaturas ao invés de usar suas assinaturas reais, praticando para manter a consistência espacial e temporal. Na primeira sessão, os colaboradores forneceram 10 assinaturas genuínas, com prática permitida antes da coleta. Na segunda sessão, que ocorreu ao menos uma semana depois da primeira, os participantes forneceram outras 10 assinaturas genuínas

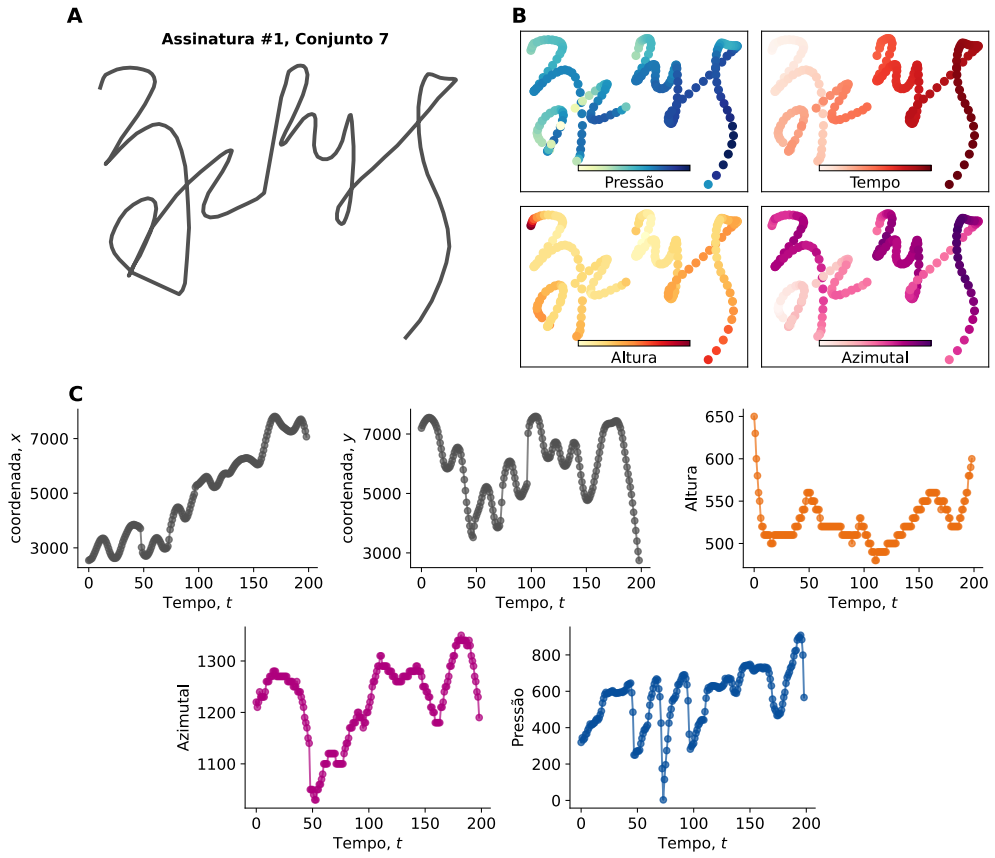

Figura 2.7: Exemplo de assinatura e das propriedades medidas durante a escrita. (A) Exemplo de assinatura original correspondente ao primeiro conjunto de número 7. A linha foi criada interpolando os pontos de acordo com a evolução temporal. Esse processo foi feito apenas para fins ilustrativos, pois os dados não são contínuos. (B) Ilustração das propriedades coletadas durante a assinatura. Os pontos mostram os valores associados à pressão, tempo, ângulo de altitude e ângulo azimutal, respectivamente, em azul, vermelho, amarelo e rosa. As tonalidades indicam a intensidade relativa dos valores. (C) Propriedades representadas como séries temporais, incluindo as coordenadas x e y como função do tempo t.

e quatro falsificações habilidosas das assinaturas de outros cinco participantes, usando um visualizador para ver e praticar as assinaturas genuínas antes de falsificá-las.

Nos dados, cada assinatura é representada como um conjunto de pontos, os quais possuem coordenadas x e y como função do tempo ao longo do processo de escrita. Além disso, os dados também contemplam a pressão, os valores do ângulo de altitude e os valores do ângulo azimutal durante a escrita, propriedades que, em conjunto com o tempo, denotamos como z_i. Mantivemos a nomenclatura usada na base de dados [54], de modo que o ângulo de altitude se refere ao ângulo entre a caneta e o plano da escrita, representando o quão vertical a caneta está. Assim, ao considerar as assinaturas como pontos espalhados no plano da escrita, estamos lidando com dados irregulares. Um exemplo de uma assinatura genuína é mostrado na Figura 2.7A, na qual interpolamos os pontos para melhor visualização. Por

outro lado, a Figura 2.7B representa as propriedades em conjunto com a estrutura espacial, enquanto a Figura 2.7C mostra as mesmas propriedades, bem como as coordenadas x e y como séries temporais.

Diferentemente das aplicações anteriores, vamos usar todos os valores da distribuição de probabilidade ordinal $P = \{\Pi_j\}_{j=1,\ldots,d!}$ no lugar de usar apenas a entropia de permutação k-nn. Além disso, vamos comparar nossos resultados com aqueles obtidos ao usar a mesma distribuição ordinal calculada a partir das representações do dado como séries temporais. Nossa ideia é verificar se existe perda significativa de informação ao ignorar a estrutura espacial dos dados quando seguimos o procedimento usual de Bandt-Pompe no lugar de analisar os padrões extraídos como nossa abordagem que explicitamente considera a estrutura espacial dos dados.

De maneira mais específica, utilizamos nossa abordagem considerando as coordenadas das assinaturas como os vetores $\vec{r}_i = (x_i, y_i)$ e as propriedades pressão, tempo, ângulo de altitude e ângulo azimutal como valores z_i. Assim, para cada propriedade, uma assinatura é mapeada em um espaço de $d!$ dimensões, sendo cada dimensão correspondente a probabilidade de um dos $d!$ padrões ordinais. Para o caso do procedimento usual de Bandt-Pompe, consideramos as séries temporais das propriedades coletadas e estimamos a distribuição ordinal para cada uma. Sendo assim, também nesse caso as assinaturas são mapeadas em um espaço de $d!$ dimensões; porém, os padrões ordinais dessa abordagem se referem aos padrões extraídos apenas das séries temporais.

Em uma tentativa de visualizar da estrutura do espaço de padrões ordinais obtidas a partir de nosso método, utilizamos o algoritmo de redução de dimensionalidade UMAP (*uniform manifold approximation and projection*) [34, 35]. De maneira resumida, podemos considerar que o UMAP cria uma representação vetorial de baixa dimensão de uma espaço de alta dimensão de modo a tentar preservar aspectos topológicos da estrutura de alta dimensão. A Figura 2.8 mostra uma projeção UMAP da distribuição ordinal dos padrões obtidos ao considerar o tempo como variável z_i, dimensão de *embedding* $d = 6$ e $k = 30$ vizinhos.

Essa projeção indica que que nem todas as assinaturas estão separadas uniformemente no espaço, permitindo-nos delinear três comportamentos principais, os quais denominamos por classes de conjuntos. A primeira classe consiste em assinaturas fortemente agrupadas em regiões distintas, principalmente na borda à direita. Essas assinaturas ocupam regiões únicas no espaço dos padrões, tais como o conjunto 14, representado por círculos na cor verde escura que se apresenta separado do corpo principal da projeção. A segunda classe representa um regime intermediário de unicidade, como ilustrado pelo conjunto 12 que está representado por quadrados na cor azul escura. Esse conjunto tem um grupo de assinaturas na parte inferior e outro disperso na parte superior à direita; porém, ainda assim, aqueles dentro de cada grupo estão próximos entre si. Isso indica que há regiões desconectadas no espaço dos padrões, mas dentro de cada região as assinaturas de uma mesma pessoa

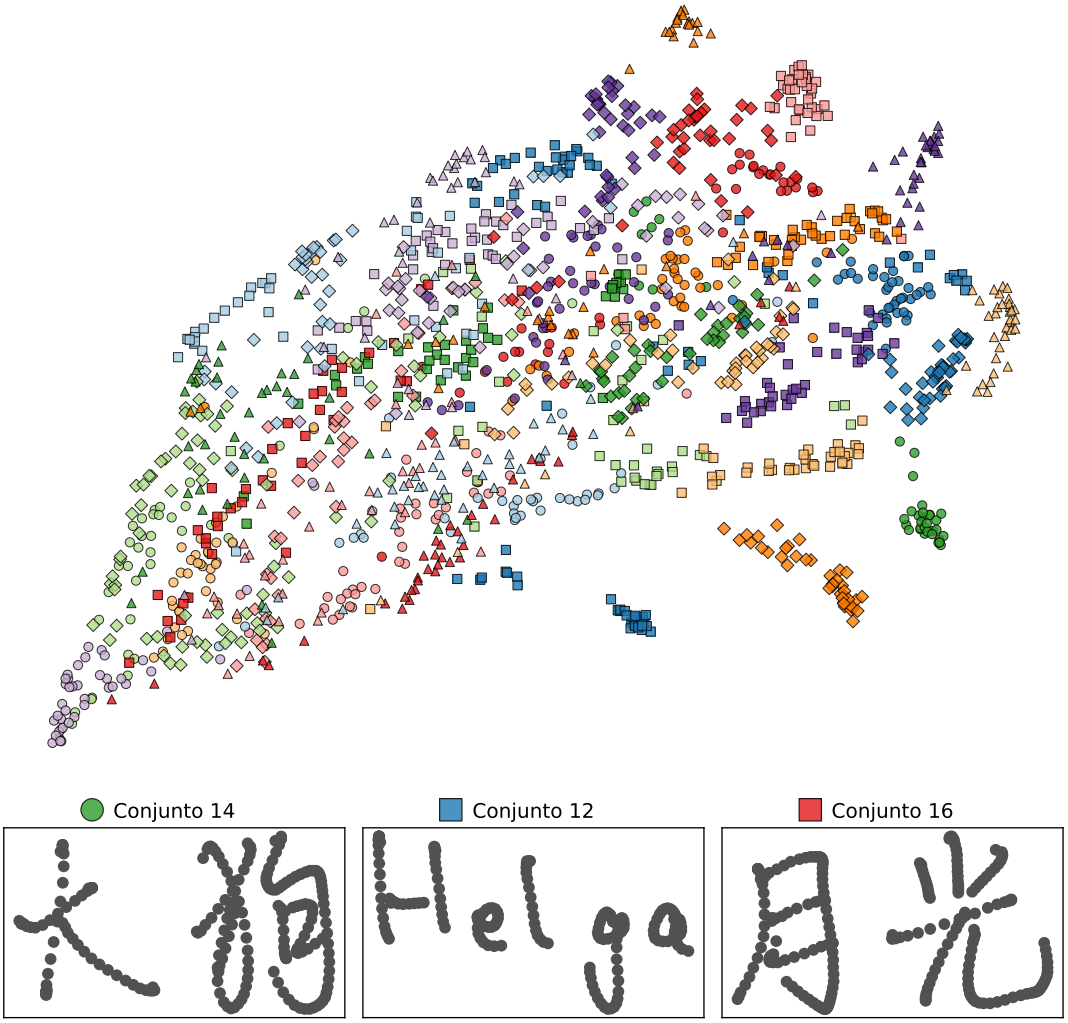

Figura 2.8: Projeção UMAP das distribuições de padrões ordinais calculadas a partir de nosso método. Os padrões ordinais foram extraídos usando o tempo como valores z_i, dimensão de *embedding* $d = 6$ e $k = 30$ vizinhos. Desse modo, as assinaturas em nosso conjunto de dados são representadas como vetores de $6! = 720$ dimensões. Nessa figura, cada combinação de marcador e cor específica corresponde a um conjunto distinto de assinaturas, sem distinção entre genuínas e forjadas. Abaixo da projeção UMAP, destacamos três exemplos ilustrativos de assinaturas.

permanecem próximas. A terceira classe, ilustrada pelo conjunto 16 e representada por quadrados na cor vermelho escura, carece de padrões distintos e está espalhada por todo o corpo principal da projeção. Este espalhamento generalizado indica que essas assinaturas distintas têm distribuições de padrões similares entre si. Além disso, é interessante observar que não há uma separação clara entre assinaturas chinesas e ocidentais, sugerindo uma certa universalidade nos padrões ordinais com relação ao idioma das assinaturas.

Para comparar sistematicamente a eficácia preditiva da nossa abordagem com o procedimento usual de Bandt-Pompe, propomos algumas tarefas de classificação utilizando algoritmos de aprendizado de máquina. Nesse caso, empregamos o algoritmo *XGBoost* [55] (*extreme gradient boosting*, veja o Apêndice A), que é uma implementação avançada do mé-

todo *boosting* de gradiente, projetada para alta eficiência computacional e acurácia. Esse método constrói uma série de árvores de decisão sequencialmente, de modo que cada nova árvore corrige os erros das anteriores por meio da otimização de uma função de perda. Para isso, alocamos 20% dos dados para teste e usamos o restante para treinamento. Treinamos o modelo *XGBoost* utilizando as distribuições ordinais extraídas das duas abordagens. No entanto, o uso direto das probabilidades nesse espaço de alta dimensão gera modelos computacionalmente ineficientes. Para resolver esse problema, reduzimos a dimensão da distribuição ordinal por meio da análise de componentes principais ou PCA na sigla inglesa para *principal component analysis*. Esse método de redução de dimensionalidade transforma as probabilidades ordinais em um conjunto de componentes linearmente não correlacionados. Essas componentes são classificados pela quantidade de variância que explicam dos preditores originais, retendo as informações mais significativas ao mesmo tempo que reduzem o número de dimensões do espaço. A depender da dimensão de *embedding*, os padrões extraídos de ambos os métodos foram reduzidos para 5 (quando $d = 3$) ou 10 (quando $d > 3$) componentes PCA.

De maneira mais específica, propomos quatro tarefas de aprendizado de máquina: *i)* classificação de todo o conjunto de dados de modo a distinguir entre diferentes conjuntos de assinaturas; *ii)* classificação usando apenas assinaturas genuínas; *iii)* classificação usando apenas assinaturas falsificadas; e *iv)* verificação se uma assinatura é genuína ou falsificada. Nessa última tarefa estamos interessados em distinguir entre assinaturas genuínas e falsificadas e entre diferentes conjuntos de assinaturas ao mesmo tempo, ou seja, o modelo de aprendizado de máquina tenta prever um rótulo categórico da assinatura e a autenticidade.

Na Figura 2.9, mostramos a comparação de ambos os métodos utilizando apenas espaço e tempo. No caso do procedimento usual de Bandt-Pompe, os padrões extraídos das coordenadas x e y foram usados em conjunto, formando vetores $2 \times d!$. Já no caso da nossa abordagem, o tempo foi considerado valores z_i. Notamos que nosso método resultou em uma acurácia significativamente maior ao usar parâmetros específicos. Na tarefa de verificação, com $d = 6$ e $k = 30$, temos uma acurácia de 60%. Notavelmente, a dimensão de *embedding* $d = 3$ resultou em acurácia mais baixa, enquanto dimensões mais altas forneceram resultados melhores. Esse resultado indica que os padrões extraídos com $d = 3$ não caracterizam as assinaturas de maneira única e que os padrões únicos de um conjunto de assinaturas surgem apenas em dimensões *embedding* maiores. Nesse sentido, o desempenho similar para $d = 5$ ou $d = 6$ sugere que não há mais informações que possam ser extraídas usando valores maiores de d. Além disso, o número de vizinhos k também afeta significativamente a acurácia da previsão em todas as tarefas, sendo as acurácias significativamente maiores com $k = 30$ do que com $k = 3$. Dado que esse parâmetro permite que o caminhante amostre estruturas espaciais mais globais ou mais locais, podemos concluir que as estruturas espaciais das assinaturas possuem mais detalhes distintivos em grandes escalas espaciais. De fato, quando

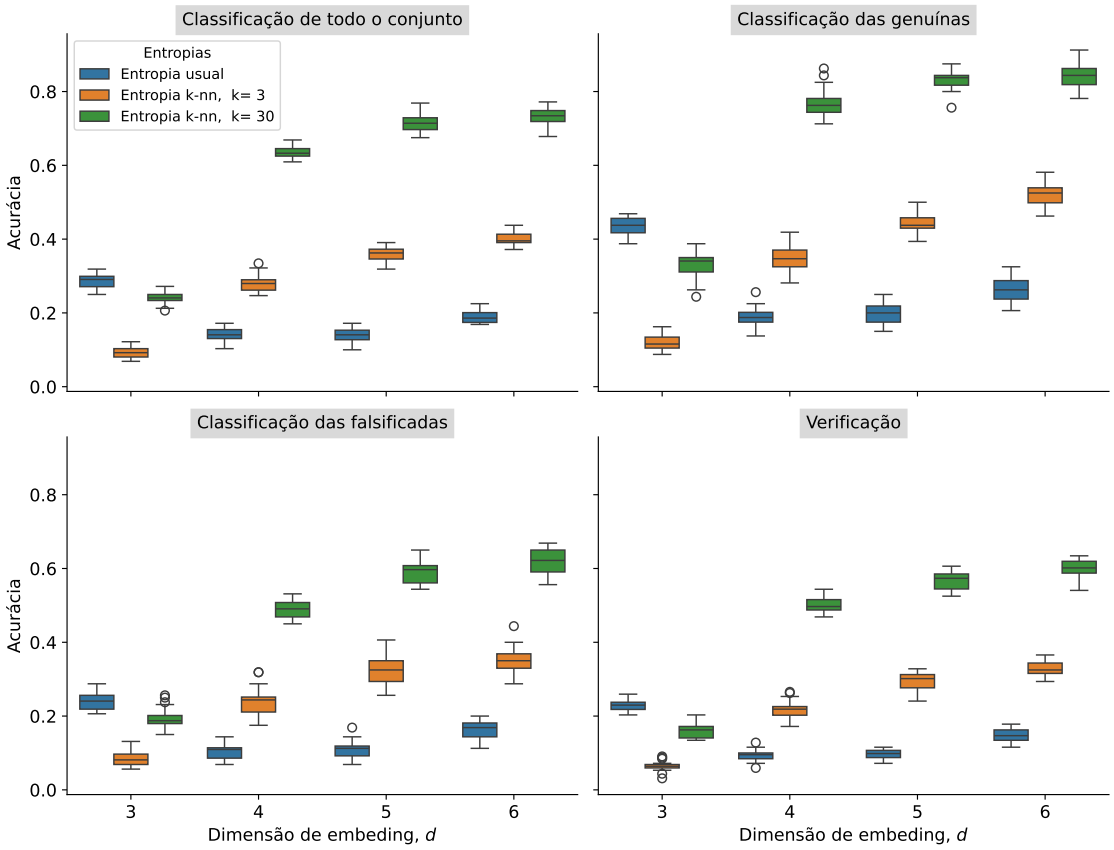

Figura 2.9: Comparação das análises de padrões ordinais baseadas em nosso procedimento e na abordagem usual de Bandt-Pompe para classificar assinaturas. O painel superior à esquerda mostra a classificação considerando todo o conjunto de dados. O painel superior à direita mostra a classificação das assinaturas genuínas. O painel inferior à esquerda mostra a classificação das assinaturas falsificadas. O painel inferior direita mostra a tarefa de verificação das assinaturas genuínas e falsificadas. Em todos os painéis, a cor azul representa o procedimento usual de Bandt-Pompe, enquanto as cores laranja e verde representam nossa abordagem com $k = 3$ e $k = 30$, respectivamente. Além disso, nessa representação de diagrama de caixa, as caixas mostram o primeiro quartil da acurácia enquanto as linhas horizontais representam os valores médios. Os bigodes se estendem para mostrar o restante da distribuição, exceto por *outliers*, que são representados por círculos vazios. Ao longo do eixo horizontal mostramos os resultados para diferentes valores da dimensão de *embedding*.

analisadas em escala pequena de espaço, as assinaturas exibem predominantemente estruturas espaciais semelhantes a linhas, que são bem capturadas ao usar $k = 3$ pois os pontos são conectados em uma ordem quase sequencial. Ao usar mais vizinhos, capturamos estruturas mais complexas que podem surgir onde as linhas se cruzam ou estão próximas, mantendo ainda a estrutura sequencial como um subgrafo. Notamos também que o procedimento usual de Bandt-Pompe obteve um desempenho pior em todas as tarefas independentemente da dimensão de *embedding*, o que pode ser associado à sua incapacidade de lidar adequadamente com a estrutura espacial dos dados.

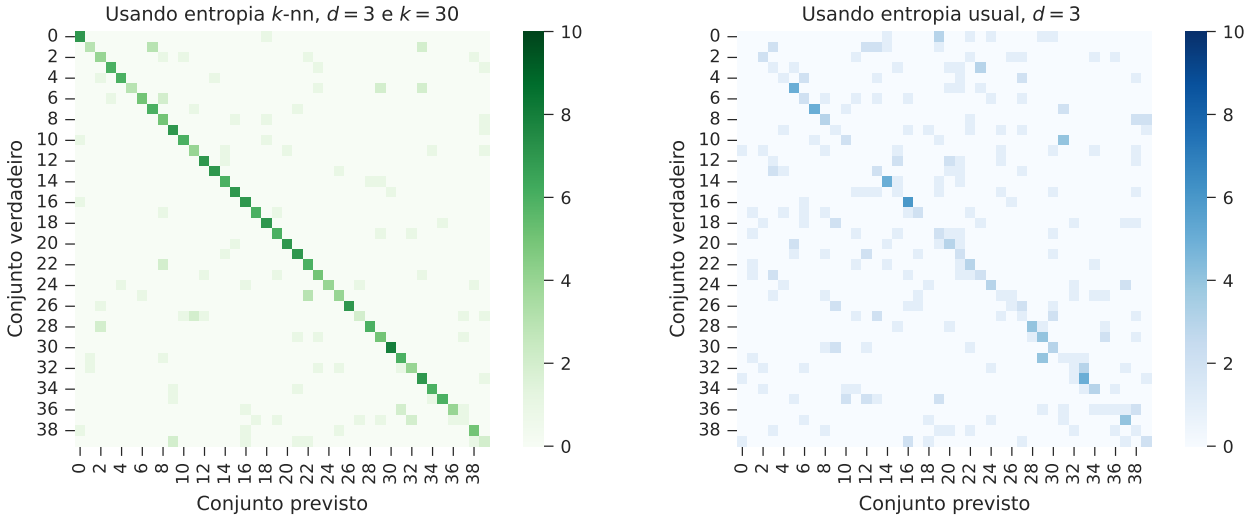

Figura 2.10: Comparação das matrizes de confusão. A tarefa compreende a classificação das assinaturas usando todo o conjunto de dados. A matriz à esquerda em verde é baseada no nosso procedimento de cálculo da distribuição ordinal, enquanto a matriz à direita em azul é baseada no procedimento usual de Bandt-Pompe. Nos dois casos, as tonalidades das cores indicam a fração de classificações corretas.

Comparamos também as matrizes de confusão para a tarefa de verificação usando os parâmetros que resultaram nas maiores acurácias das duas abordagens, como mostrado na Figura 2.10. Para o nosso método, a matriz de confusão exibe uma estrutura diagonal bem definida. Em contraste, a matriz para o método usual de Bandt-Pompe mostra apenas uma diagonal tênue, refletindo um desempenho inferior. Dado que o conjunto de dados compreende assinaturas chinesas e ocidentais, não observamos diferença significativa na acurácia da previsão entre os dois tipos de escrita.

Em ambas as abordagens, há uma diferença clara na classificação dos conjuntos de assinaturas genuínas e falsificadas. Além disso, a acurácia da classificação das assinaturas genuínas sozinhas é maior do que a do conjunto de dados completo, indicando que a introdução de assinaturas falsificadas aumenta a confusão no modelo. Isso também indica que os conjuntos de assinaturas falsificadas estão mais relacionados entre si, enquanto as assinaturas dentro de cada conjunto são mais diferentes entre si. Sabemos que assinaturas falsificadas são escritas planejadas, enquanto assinaturas genuínas estão associadas a movimentos naturais [56, 57]. Portanto, a maneira natural de escrever uma assinatura parece ser mais única, no sentido de ser representada por padrões ordinais de tempo, do que ao tentar copiar uma. Também existe a possibilidade de que os padrões caracterizem os indivíduos que escreveram as assinaturas, dado que todas as 20 assinaturas genuínas em cada conjunto foram escritas pela mesma pessoa, enquanto as assinaturas em cada um dos conjuntos falsificados foram escritas por pessoas diferentes [54].

Para testar e comparar se as outras propriedades são tão representativas quanto o tempo, extraímos padrões ordinais usando as séries temporais de cada uma. Em contraste, nosso

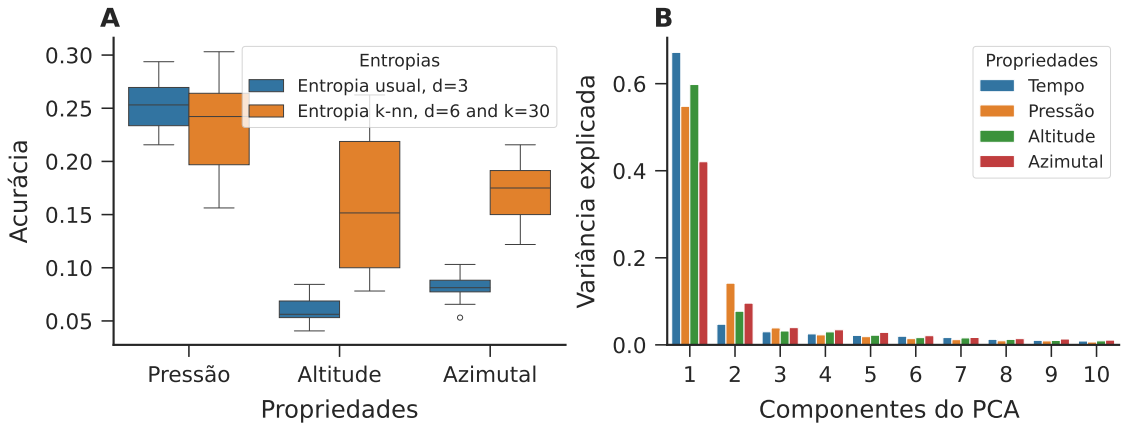

Figura 2.11: Acurácia nas tarefas de verificação usando padrões das outras propriedades do dado e a variância explicada de cada componente do PCA. (A) Comparação da acurácia na tarefa de verificação para as propriedades pressão, ângulo de altitude e ângulo azimutal das assinaturas. Em azul mostramos o desempenho da abordagem usual de Bandt-Pompe com dimensão de *embedding* $d = 3$ (que foi a melhor). Em laranja mostramos o desempenho da nossa abordagem com $d = 6$ e $k = 30$. (B) Variância explicada dos componentes PCA para o tempo, a pressão, o ângulo de altitude e o ângulo azimutal, conforme indicado na legenda.

método extraiu padrões dos dados usando a estrutura espacial e os valores z_i associados a cada ponto, permitindo aproveitar simultaneamente a estrutura dos valores de $\{z_i\}$ e a estrutura espacial das propriedades. A Figura 2.11A demonstra que a pressão teve o melhor desempenho, superando até mesmo o tempo no caso do procedimento usual. Os valores do ângulo de altitude e do ângulo azimutal levam a acurácias menores, sugerindo que a posição da caneta de uma pessoa varia muito dentro de seu conjunto de assinaturas o torna essas variáveis menos importantes para classificar as assinaturas.

Por fim, para confirmar que o número de componentes PCA usados é suficiente, mostramos a variância explicada de cada componente para a tarefa de verificação feita usando as duas abordagens na Figura 2.11B, na qual observamos que as duas primeiras componentes já explicam quase a totalidade da variância dos dados.

CAPÍTULO 3

Caracterização de imagens

Na literatura, a primeira tentativa de aplicar a entropia de permutação a imagens foi feita calculando padrões ordinais a partir de partições bidimensionais de tamanho $d_x \times d_y$ (as dimensões *embedding*) em matrizes regulares representando intensidades de pixels em uma imagem [29]. No contexto da nossa abordagem, uma outra possibilidade envolve considerar imagens como dados dispostos em uma estrutura espacial regular que corresponde aos pixels arranjados nas posições de uma rede quadrada. Para tal, podemos considerar as coordenadas espaciais como $\vec{r}_t = (i, j)$ e associar as intensidades dos pixels x_{ij} aos valores z_t, com $t = 1, \ldots, N$ enumerando todos os $N = N_x N_y$ pixels em uma imagem e $i = 1, \ldots, N_x$ e $j = 1, \ldots, N_y$ servindo de índices para as linhas e colunas da representação matricial da imagem.

Uma das possíveis vantagens de calcular a entropia de permutação k-nn nesse tipo de dado é que nossa abordagem permite a extração de padrões ordinais mais gerais de uma matriz de imagem $\{x_{ij}\}_{i=1,\ldots,N_x}^{j=1,\ldots,N_y}$. Além disso, ao variar o número de vizinhos k, criamos estruturas de grafos de vizinhos mais próximos que permitem acessar padrões ordinais em diferentes escalas espaciais, como pode ser visto na Figura 3.1. Por exemplo, ao conectar os primeiros vizinhos dos pixels de uma imagem ($k = 4$), obtemos um grafo que possibilita a extração de padrões ao longo de trajetórias limitadas pela ordem dos pixels adjacentes. Ao conectar os primeiros e segundos vizinhos ($k = 8$), temos a adição de arestas entre as diagonais dos pixels, produzindo amostras de pixels ao longo das direções vertical, horizontal e diagonal. Seguindo esse procedimento, podemos adicionar os terceiros vizinhos ($k = 12$) com o grafo gerado conectando aos segundos vizinhos nas direções horizontal e diagonal. Nessa representação, os segundos vizinhos são vistos como tendo a mesma distância que os primeiros vizinhos e o caminhante terá a mesma probabilidade de "pular" para um pixel ou mover-se para um pixel

adjacente. Isso permite que os caminhantes explorem não apenas padrões sequenciais, mas também padrões segmentados. Isso significa que a entropia pode extrair padrões presentes em diferentes escalas na imagem. À medida que o número de vizinhos aumenta, a possibilidade de uma maior heterogeneidade espacial nas amostras também aumenta.

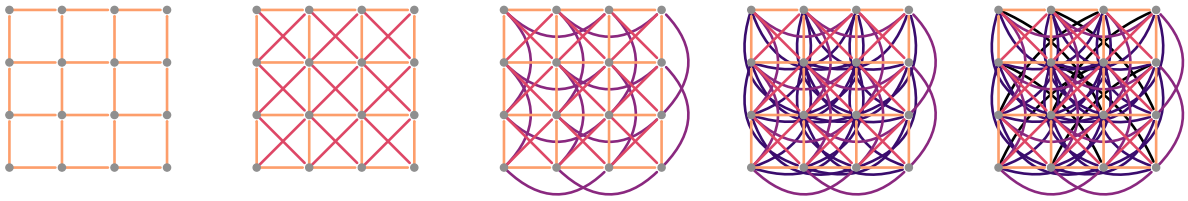

Figura 3.1: Grafos em redes regulares. Ilustrações de grafos regulares formados ao conectar pixels em uma imagem 4×4 aos seus primeiros ($k = 4$), segundos ($k = 8$), terceiros ($k = 12$), quartos ($k = 16$) e quintos ($k = 20$) vizinhos.

Para analisar sistematicamente essas possibilidades, neste capítulo, empregamos a entropia de permutação k-nn para caracterizar simulações de texturas de cristais líquidos colestérico com o intuito de prever uma propriedade física desse tipo de material usando métodos de aprendizado de máquina. Além disso, comparamos nossos resultados com aqueles obtidos ao usar a entropia de permutação usual de dados em duas dimensões.

3.1 Texturas de cristais líquidos colestéricos

Nossa investigação tem como objetivo inicial examinar texturas de cristais líquidos colestéricos com diferentes comprimentos de passo (uma propriedade física desse material que apresentaremos a seguir) que foram previamente estudados usando a entropia de permutação usual [7]. Nosso objetivo é realizar uma análise similar utilizando a entropia de permutação k-nn. Pretendemos demonstrar que os valores da entropia são distintos para cada comprimento de passo e, usando aprendizado de máquina, iremos verificar a capacidade prever o valor do comprimento de passo de cada textura com base no valor da entropia.

O estado líquido cristalino é uma mesófase entre o sólido cristalino e o líquido isotrópico possuindo propriedades de birrefringência de um cristal e podendo fluir como um líquido [58]. A ordem em um cristal líquido pode ser pensada como o quão alinhado seus componentes estão em relação a um eixo preferencial. Cada molécula tem sua direção representada por um vetor $\vec{a}$. O vetor diretor $\vec{n}$ de um cristal líquido é a média dos eixos preferenciais das moléculas. No entanto as direções $+\vec{a}$ e $-\vec{a}$ são equivalentes, sendo assim, o vetor diretor, de certa forma, não tem sentido e apenas representa uma direção. Tal média é tomada em um volume grande comparado às dimensões da molécula, mas pequeno o suficiente quando comparado ao comprimento das deformações do vetor diretor, sendo possível associar uma função contínua ao vetor diretor, $\vec{n}(\vec{r})$, com $\vec{r}$ representando uma posição na amostra.

Nesse contexto, o grau de alinhamento pode ser medido por meio da quantidade S, denominada parâmetro de ordem uniaxial e definida como

$$S = \frac{1}{2}\left\langle 3\cos^2\theta - 1\right\rangle = 2\pi\int_0^\pi P_2[\cos(\theta)]f(\theta)\sin(\theta)d\theta \tag{3.1}$$

em que $P_2[\cos(\theta)]$ é o segundo polinômio de Legendre e θ é o ângulo formado entre o eixo principal da molécula e o diretor. Além desse direcionamento principal, as moléculas também podem se organizar de tal maneira a criar uma segunda direção privilegiada representada pelo vetor codiretor $\vec{l}$ que é perpendicular a $\vec{n}$. Assim como o parâmetro de ordem uniaxial, é possível criar um grau de alinhamento com relação ao codiretor definido por

$$P = \left[\langle P_2(\vec{a}.\vec{l})\rangle - \langle P_2(\vec{a}.\vec{m})\rangle\right], \tag{3.2}$$

com $\vec{m} = \vec{n}\times\vec{l}$ [59]. Tal parâmetro está definido no intervalo $[-\frac{3}{2}, \frac{3}{2}]$, sendo que, $P = 0$ é o ordenamento uniaxial e $P = \frac{3}{2}$ representa um ordenamento biaxial ao longo do codiretor. Os possíveis valores de P estão limitados por S, ou seja, $-(1-S) \leq P \leq (1-S)$, de tal modo que o ordenamento biaxial perfeito só é possível quando $S = -\frac{1}{2}$.

Também é possível definir um tensor de ordem 2 que englobe os dois parâmetros de ordem

$$Q_{ij} = \frac{1}{2}S(3n_in_j - \delta_{ij}) + \frac{1}{2}P(l_il_j - m_im_j), \tag{3.3}$$

no qual i, j e k variam de um a três e representam a base canônica. Sendo assim, os índices dos versores representam a decomposição do vetor na direção da respectiva coordenada. Esse parâmetro de ordem tensorial é suficiente para descrever cristais líquidos nemáticos e colestéricos. A diferença entre os dois é que o colestérico não possui simetria quiral, ou seja, ele é diferente da sua imagem espelhada. Os colestéricos possuem uma estrutura em que a orientação das moléculas gira em torno de um eixo, formando uma hélice. Dado o eixo helicoidal, à medida que se percorre esse eixo, o vetor diretor $\vec{n}$ gira e o período espacial para que $\vec{n}$ dê uma volta completa é igual a $l = \frac{\rho}{2}$, com ρ sendo conhecido como comprimento do passo do colestérico. Devido a essa estrutura periódica, os colestéricos produzem reflexões de Bragg, sendo assim, o passo também pode ser entendido como o comprimento de onda da luz refletida de Bragg.

No modelo fenomenológico de Landau-de Gennes, a energia livre deve ser expandida em potências que são combinações invariantes dos elementos de Q_{ij}, isto é,

$$\begin{aligned} f_{LdG} = {} & \frac{A}{2}Q_{ij}Q_{ji} + \frac{B}{3}Q_{ij}Q_{jk}Q_{ki} + \frac{C}{4}(Q_{ij}Q_{ji})^2 + \frac{L_1}{2}\frac{\partial Q_{ij}}{\partial x_k}\frac{\partial Q_{ij}}{\partial x_k} + \\ & \frac{L_2}{2}\frac{\partial Q_{ij}}{\partial x_j}\frac{\partial Q_{ik}}{\partial x_k} + \frac{L_3}{2}Q_{ij}\frac{\partial Q_{kl}}{\partial x_i}\frac{\partial Q_{kl}}{\partial x_j} + \frac{4\pi}{\rho}L_q\epsilon_{ikl}Q_{ij}\frac{\partial Q_{ij}}{\partial x_k}. \end{aligned} \tag{3.4}$$

Os três primeiros termos são necessários para descrever transição de fase, sendo que A, B e C são parâmetros termodinâmicos. Por exemplo, uma transição de fase é descrita por uma mudança no sinal de $A(T)$. Os outros termos quantificam uma penalidade energética devido às distorções elásticas, sendo que L_1, L_2, L_3 e L_q são constantes elásticas. Em especial, ρ o comprimento do passo colestérico. A evolução temporal do parâmetro de ordem tensorial é dado pela equação

$$\Gamma \frac{\partial Q_{ij}}{\partial t} = \frac{\partial f_{LdG}}{\partial Q_{ij}} - \frac{d}{dx_k} \frac{\partial}{\partial f_{LdG} \frac{Q_{ij}}{\partial x_k}}, \tag{3.5}$$

na qual usamos a notação de soma de Einstein para índices repetidos. Além disso, Γ é a viscosidade rotacional do cristal líquido e t é o tempo.

Para simular texturas colestéricas, podemos resolver numericamente a evolução temporal das componentes do parâmetro de ordem via a Equação 3.5. Nos reportamos a Sigaki *et al.* [7] para mais detalhes sobre essas simulações e a obtenção das texturas. A Figura 3.2 mostra exemplos dessas texturas para seis diferentes comprimentos de passo, nas quais os pixels representam a intensidade da luz transmitida através das amostras. Para nossas análises, utilizamos 355 texturas de tamanho 150×150 para cada comprimento de passo ρ entre $\rho = 15$ nm a $\rho = 29$ nm em passos de 2 nm.

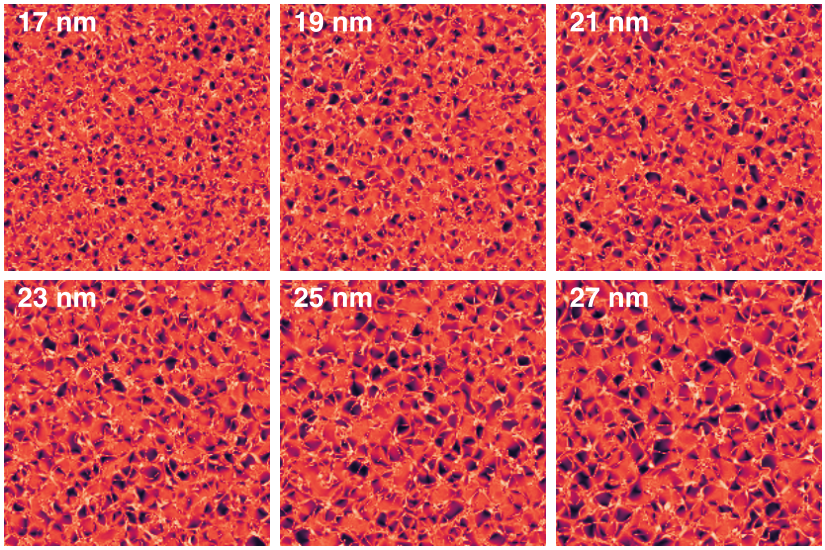

Figura 3.2: Texturas de cristais líquidos colestéricos. Exemplos de texturas ópticas de cristais líquidos colestéricos com diferentes passos ρ. A intensidade da luz transmitida é representada pelos tons de laranja, sendo tons mais claros indicativo de maior intensidade e o tons mais escuros de menor intensidade da luz.

Calculamos a entropia de permutação k-nn S para todas as texturas em nosso conjunto de dados, utilizando números de vizinhos $k = 4, 8, \ldots, 24$ (correspondente aos primeiros até os sextos vizinhos) e dimensões *embedding* $d = 4, 5$. Subsequentemente, avaliamos a relação entre os valores médios de S e os comprimentos de passo ρ, comparando esses resultados com os da entropia de permutação bidimensional usual com $d_x = d_y = 2$.

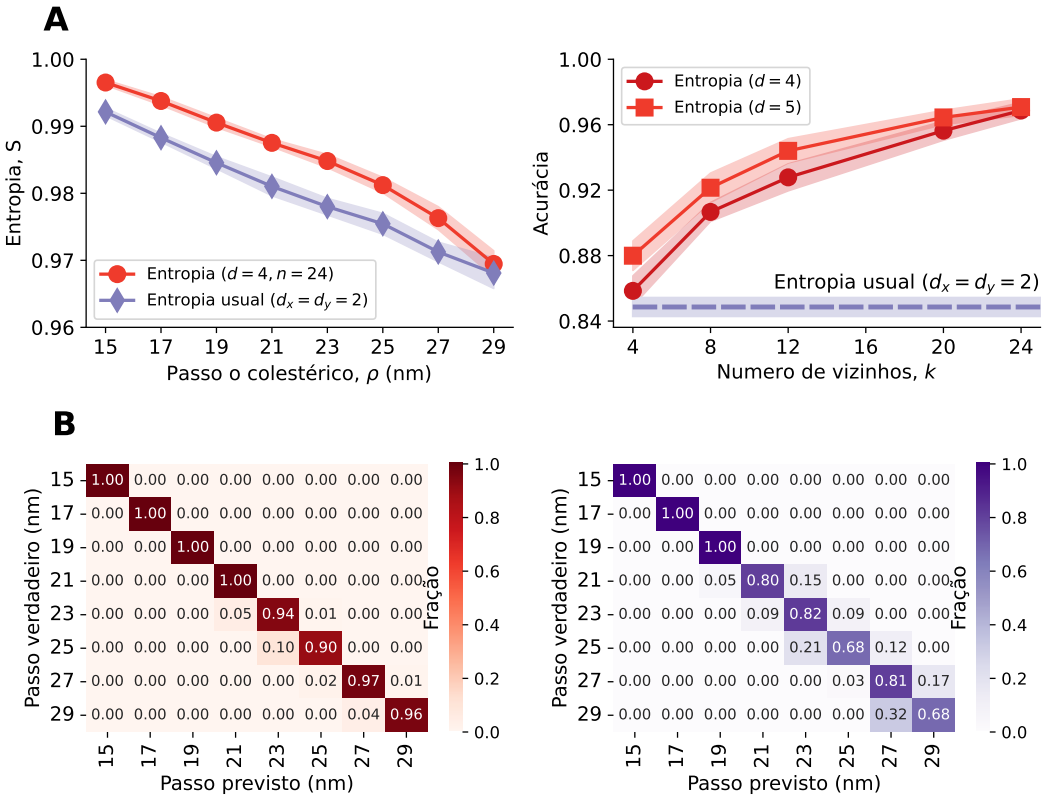

Figura 3.3: Comparação entre a entropia de permutação k-nn e a entropia usual em imagens. (A) À esquerda mostramos os valores das entropia em função do passo ρ. Em vermelho representamos a entropia de permutação dos k-primeiros vizinhos e em azul a entropia usual. À direita mostramos a análise da acurácia em tarefas de classificação com o objetivo de prever o passo ρ usando as duas entropias com diferentes valores da dimensão de *embedding* $d = \{4, 5\}$ e diferentes valores de primeiros vizinhos $k = \{4, 8, \ldots, 24\}$. (B) Exemplo de matrizes de confusão resultantes da aplicação do algoritmo de aprendizado para prever o valor de ρ usando a entropia de permutação k-nn com $d = 4$ e $k = 24$ (tons de vermelho) e usando a entropia usual (tons de roxo).

Como mostrado no gráfico à esquerda da Figura 3.3A, ambas as medidas de entropia (para $k = 24$ e $d = 4$) exibem uma tendência decrescente semelhante com o comprimento do passo. As áreas sombreadas representam um desvio padrão de cada quantificador entre as texturas para cada passo. Um exame detalhado dessas curvas revela que a variabilidade na entropia de permutação k-nn é menor do que a variabilidade da entropia de permutação usual. Para verificar se essa diferença é suficiente para melhora a eficácia da classificação com o uso da entropia de permutação k-nn, aplicamos a mesma metodologia de aprendizado de máquina de nossas investigações anteriores para categorizar texturas colestéricas utilizando cada medida de entropia como uma característica preditiva. Os resultados no gráfico à direita da Figura 3.3A mostram a acurácia média dos classificadores treinados com valores da entropia de permutação k-nn usando um número crescente de vizinhos k, comparados à acurácia média alcançada com a entropia de permutação usual para. Observamos que usar a

entropia de permutação k-nn melhora marginalmente a acurácia quando $k = 4$ (86% versus 85% para a entropia convencional). No entanto, à medida que o número de vizinhos usados para calcular a entropia de permutação k-nn aumenta, a disparidade entre os dois métodos se torna mais pronunciada. Notavelmente, a acurácia atinge 96% para $k = 24$, um desempenho comparável a métodos mais sofisticados baseados em redes neurais convolucionais profundas [60].

A Figura 3.3B compara matrizes de confusão típicas dos classificadores treinados com ambas as medidas de entropia, destacando que a entropia de permutação k-nn melhora principalmente a acurácia da classificação para valores altos do passo. Portanto, os padrões ordinais originados da estratégia de amostragem introduzida por nossa abordagem de fato fornecem mais informações sobre a estrutura da imagem, o que se traduz em um desempenho de classificação aprimorado.

O comprimento do passo é uma propriedade fundamental de cristais líquidos colestéricos, pois além de determinar a periodicidade da estrutura helicoidal, ele influencia diretamente a aparência óptica e propriedades físicas do material. O passo também pode definir se a textura observada é homeotrópica, de impressão digital ou focal-cônica [58]. Além disso, o passo afeta a formação e a estabilidade de defeitos topológicos, a resposta a estímulos externos como campos elétricos e magnéticos, e a capacidade de refletir luz em comprimentos de onda específicos, crucial para aplicações em mostradores, sensores ópticos e materiais fotônicos [61]. Mais ainda, existe aplicação prática em tentar obter o passo a partir das texturas, pois quando visto em um microscópio óptico essa propriedade só é facilmente determinada quando o eixo helicoidal está perpendicular à direção de visualização, o que nem sempre ocorre. Assim existe a perspectiva do uso da entropia k-nn como estimador do passo em situações experimentais.

As análises realizadas até agora comprovam que a entropia k-nn é perfeitamente capaz de analisar imagens, possibilitando uma série de aplicações e também revisitar outros problemas analisados com a entropia usual como a classificação de obras de arte [9]. Outra possibilidade é usar nossa entropia em dados volumétricos, seguindo o mesmo princípio que foi usado nas imagens. Os dados de um campo volumétrico podem ser imaginados como distribuídos em uma rede cúbica regular, sendo um processo pontual em três dimensões no qual os pontos seriam representados por um vetores no espaço com valores associados. Dentre outras coisas, seria possível analisar o *bulk* em transições de fase em cristais líquidos, nas quais os valores z_i poderiam representar a ordem local das moléculas. De modo geral, podemos usar nossa abordagem para analisar campos de densidade, temperatura, pressão, entre outras estruturas espaciais, além de dados atmosféricos cruciais para a meteorologia, climatologia e gestão ambiental.

CAPÍTULO 4

Caracterização de séries temporais

Neste capítulo, vamos mostrar que também é possível usar a entropia de permutação dos k-primeiros vizinhos para caracterizar séries temporais. Embora a entropia de permutação de Bandt e Pompe seja uma medida bem estabelecida na literatura para essa finalidade [2], nossa abordagem introduz algumas características interessantes para a análise desse tipo de dado. Conforme veremos, a entropia de permutação k-nn permite incluir informações sobre as amplitudes das séries temporais e leva em conta a existência de lacunas temporais entre as observações, informações essas que conduzem a uma maior acurácia em tarefas de aprendizado de máquina quando comparada a entropia de permutação usual.

Uma possibilidade é considerar séries temporais como um tipo de dado espalhado do plano da série no qual cada ponto tem um valor associado. Assim, se seguíssemos estritamente a lógica do primeiro passo de nosso método, o cálculo das distâncias usadas para encontrar os primeiros vizinhos deveria ser baseado apenas na distância ao longo do tempo. Os elementos adjacentes na série seriam sempre os vizinhos mais próximos, com os pontos conectados na mesma sequência em que aparecem na série. Logo, uma amostragem dos valores usando uma busca em profundidade obteria séries idênticas a secções da série temporal original, resultando em padrões ordinais iguais aos obtidos com a entropia de permutação tradicional. Nesse caso, a representação a partir do grafo não produz informação adicional e, consequentemente, ambas as medidas de entropia conduziriam a resultados praticamente idênticos. Por esse motivo, optamos por uma abordagem diferente na qual consideramos uma série temporal como um processo pontual em duas dimensões. Para isso, usamos o tempo e a dimensão do valor da série para criar um espaço bidimensional. Por exemplo, considerando uma série original $\{(x_1, t_1), (x_2, t_2), ...\}$, esse procedimento corresponde a um conjunto de pontos cuja a posição é $\vec{r}_t = (x_t, t)$ e cujo os valores z_t são os valores x_t da

série. Ao visualizar a série em um plano, os pontos mais próximos no espaço bidimensional são conectados, eliminando a necessidade de conectar os pontos na ordem exata em que eles aparecem na série. Essa abordagem garante que tanto as amplitudes quanto às lacunas de tempo sejam consideradas na construção dos grafos de k-primeiros vizinhos. Essa característica diferencia substancialmente nosso método da entropia de permutação original de Bandt e Pompe. Um aspecto sutil a ser considerado nesse procedimento é que as variáveis x e t têm naturezas diferentes, de modo que uma mudança na escala, ou unidade de medida, de uma delas não afeta necessariamente a outra. Assim, a obtenção dos k-primeiros vizinhos não é invariante sob transformações de escala ou de unidades de medida introduzindo um parâmetro adicional a ser ajustado: a escala relativa entre as coordenadas x e t. Nas aplicações apresentadas a seguir, não fizemos referência a essa escala relativa porque os resultados já se mostraram efetivos; entretanto, essa característica pode se tornar importante em outras aplicações.

4.1 Movimento browniano fracionário

Como primeiro protótipo de testes, vamos usar séries temporais geradas a partir de movimentos brownianos fracionários com diferentes expoentes de Hurst h. Essas séries são obtidas a partir da equação

$$x_t = x_{t-1} + \xi_h, \tag{4.1}$$

na qual ξ_h representa um ruído gaussiano fracionário de média zero, variância unitária e expoente de Hurst h. Esses termos de ruído são numericamente simulados usando o método de Hosking da mesma maneira que fizemos no Capítulo 2. Como pode ser visto na Figura 4.1, assim como no caso dos pontos espalhados irregularmente, essas séries apresentam padrões estruturais que variam em grau de estrutura. Para expoentes de Hurst altos, as séries apresentam correlação positiva que leva a um comportamento persistente. Para expoentes baixos, os pontos são anti-correlacionados e as séries exibem um comportamento anti-persistente. Quando o expoente é igual a $1/2$, a série se comporta como um movimento browniano usual.

Um método que estima o expoente de Hurst é a análise de flutuação destendenciada ou DFA (*detrended fluctuation analysis*) [62]. Essa técnica foi desenvolvida com o propósito inicial de analisar correlações de longo alcance em sequências de DNA. Nesse contexto, a ocorrência de purinas e pirimidinas pode ser interpretada como uma caminhada aleatória, em que a presença de uma é considerada um passo para frente e a presença da outra é um passo para trás. O DFA se tornou um dos métodos mais amplamente utilizados e confiáveis para estimar expoentes de Hurst em séries temporais [63]. A obtenção da estimativa do DFA envolve o cálculo da função de flutuação raiz-média-quadrática $F(m)$ sobre partições não sobrepostas de comprimento m provenientes de séries temporais. Isto é feito após a remoção

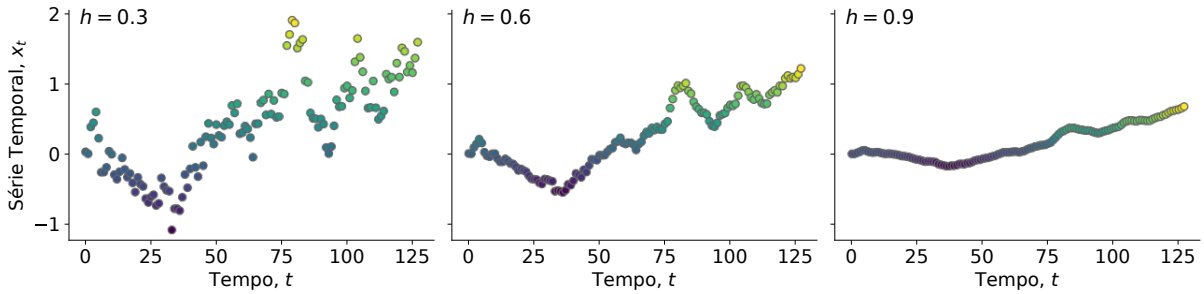

Figura 4.1: Exemplos de séries de movimentos brownianos fracionários. As figuras ilustram três séries temporais do movimento browniano fracionário gerados usando três valores do expoente de Hurst ($h = \{0.3, 0.6, 0.9\}$) e $N = 2^7 = 128$ pontos. Notamos que para o caso do expoente menor, os pontos da série são mais anti-persistentes. No caso oposto, para o expoente maior, os pontos da série são positivamente mais persistentes.

de uma tendência polinomial (aqui vamos usar polinômios lineares). Para séries temporais autossimilares, como as provenientes de movimentos brownianos fracionários, $F(m)$ apresenta uma dependência de lei de potência no comprimento da partição m, expressa como $F(m) \sim m^{h_{\text{dfa}}}$, com h_{dfa} sendo o expoente de Hurst estimado. Para estimar h_{dfa}, calculamos a inclinação da versão linearizada dessa relação de lei de potência [$\log F(m) \sim h_{\text{dfa}} \log m$] usando o método dos mínimos quadrados.

O nosso primeiro objetivo consiste em comparar o método DFA com a entropia de permutação k-nn em séries do movimento browniano fracionário após a remoção de uma fração f_r de pontos escolhidos aleatoriamente. Essas séries deixam de ser regulares no tempo uma vez que seus elementos não ocorrem em intervalos fixos de tempo. A Figura 4.2 ilustra esse tipo de série para três valores de f_r (0%, 60% e 90%). Uma particularidade interessante dos nossos dados é que o movimento Browniano fracionário é autossimilar e possui uma natureza fractal, mantendo um mesmo comportamento em diferentes escalas de tempo. Nesse sentido, a remoção de pontos tende a degradar a estrutura local dos dados e manter apenas a estrutura global, semelhante a observar os dados em uma escala maior de tempo.

Essas séries são exatamente o tipo de dado que a entropia de permutação k-nn foi desenvolvida para analisar. Da mesma maneira, o DFA também vai conseguir lidar com elas ao não levar em consideração os intervalos de tempo; em vez disso, o valor do tempo é substituído por um número sequencial que denota o índice de observação. Ambas as medidas obtiveram sucesso na caracterização dos diferentes valores de expoente de Hurst, como pode ser visto na Figura 4.3A. Essa figura ilustra a relação entre o valor do expoente estimado pelo DFA h_{dfa} e o verdadeiro expoente de Hurst h, bem como a relação entre a entropia de permutação k-nn S e o verdadeiro expoente de Hurst h, quando 45% dos pontos de dados são removidos. A associação entre h_{dfa} e h é aproximadamente linear, enquanto os valores de entropia S diminuem de maneira não linear conforme h aumenta, comportamento semelhante ao observado no movimento browniano fracionário bidimensional. Vale notar que a

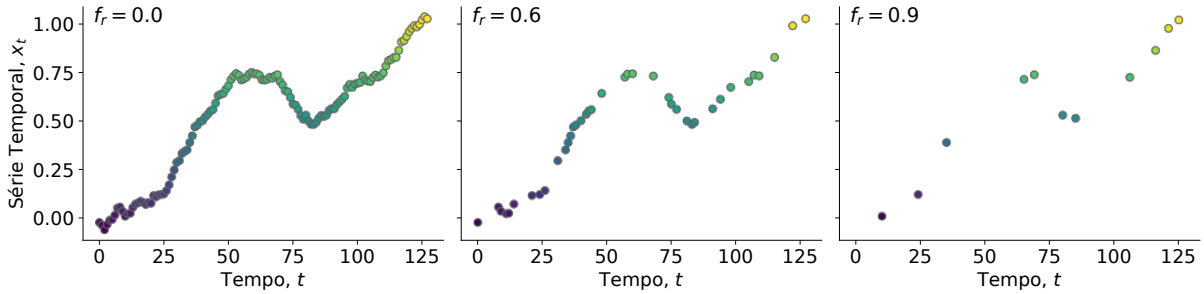

Figura 4.2: Exemplos de séries temporais de movimentos brownianos fracionários com diferentes frações de pontos removidos. A figura mostra uma mesma série temporal gerada com $h = 0.8$ e $N = 2^7 = 128$ pontos, na qual removemos diferentes frações dos pontos ($f_r \in \{0, 0.6, 0.9\}$). Devido à autossimilaridade, as séries continuam apresentando uma estrutura global parecida, apesar da remoção dos pontos.

variabilidade nos valores de S (representada por intervalos de confiança de um desvio padrão na Figura 4.3A) é significativamente menor do que a variabilidade observada para h_{dfa}, especialmente para valores mais baixos do expoente de Hurst.

Para comparar rigorosamente a capacidade das duas medidas, usamos a mesma metodologia de aprendizado de máquina de nossas investigações anteriores para classificar os expoentes de Hurst h usando os valores de entropia S (com $d = 5$) ou as estimativas derivadas do DFA h_{dfa}, em diferentes frações de remoção de dados f_r e com séries temporais de tamanhos $N = \{2^{11}, 2^{12}, 2^{13}\}$. A Figura 4.3B mostra a acurácia para cada método em função de f_r. Embora ambos os classificadores alcancem níveis altos de acurácia, o classificador baseado em valores de entropia supera o classificador baseado em DFA quando se tem um maior número de dados, especialmente quando $f_r < 0.5$, condição na qual o classificador baseado em entropia alcança uma acurácia média superior a 95%. A acurácia cai abruptamente para ambos os classificadores quando aproximadamente 80% ou mais dos pontos de dados são omitidos da série, sendo o método DFA ligeiramente mais robusto do que a entropia de permutação k-nn nessas condições.

Também comparamos as matrizes de confusão de classificadores treinados com ambas as métricas, conforme ilustra a Figura 4.4A para $f_r = 0.45$ e $d = 3$. Em geral, essas matrizes exibem um padrão diagonal pronunciado que indica a eficácia dos classificadores. No entanto, os elementos diagonais das matrizes de confusão usando os valores de S estão mais próximos de um, destacando o desempenho superior do classificador, especialmente para expoentes de Hurst maiores, nos quais os classificadores baseados em h_{dfa} apresentam maior confusão. Em relação aos parâmetros, verificamos que mudanças na dimensão de *embedding* d não afetam significativamente a acurácia da classificação, como mostrado na Figura 4.4B. Para essas classificações, usamos séries de tamanho $N = 2^{13} = 8192$.

Na literatura, há uma certa escassez de medidas para a caracterização de séries irregulares no tempo. A maioria das análises de séries temporais irregulares são realizadas a partir

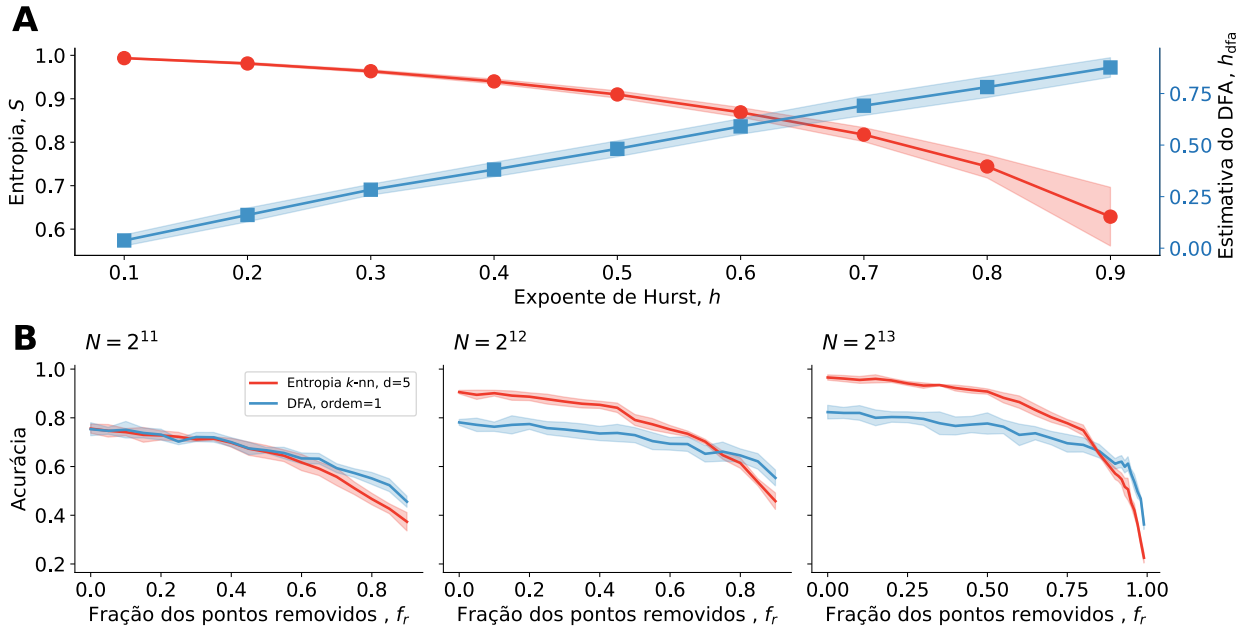

Figura 4.3: Comparação do desempenho do DFA e da entropia de permutação dos k primeiros vizinhos em séries temporais irregulares de tamanhos diferentes. (A) Relação entre a entropia de permutação dos k-primeiros vizinhos S (com $d = 5$) e o expoente de Hurst h (círculos vermelhos), junto com a associação entre o expoente de Hurst estimado pelo DFA h_{dfa} e o expoente de Hurst real h (quadrados azuis). Em ambas as curvas, os marcadores indicam valores médios estimados a partir de cem simulações distintas de movimentos brownianos fracionários, cada uma com um tamanho $N = 2^{13} = 8192$ e uma fração de pontos removidos de $f_r = 0.45$. As áreas sombreadas representam intervalos de confiança de um desvio padrão. (B) Análise da acurácia em tarefas de classificação com o objetivo de prever os expoentes de Hurst ($h \in \{0.1, 0.2, \ldots, 0.9\}$) usando a entropia de permutação dos k-vizinhos S (curva vermelha) versus estimativas do DFA (h_{dfa}, curva azul) como função da fração de pontos de dados removidos f_r. Os três painéis mostram os resultado para séries de tamanho $N = \{2^{11}, 2^{12}, 2^{13}\}$. As regiões sombreadas representam um desvio padrão dos níveis de acurácia com base em dez realizações independentes do processo de treinamento.

da recuperação das observações ausentes por meio de suavização ou interpolação [64], generalização de ferramentas de análise espectral [65] ou métodos baseados em *kernel* [66]. Nesse contexto, a entropia de permutação k-nn se destaca por ser uma medida capaz de caracterizar regimes regulares e aleatórios em séries temporais irregulares sem a necessidade de alterar os dados originais. Em termos de aplicabilidade a dados empíricos, destacamos a possibilidade de usar nossa medida para caracterizar séries provenientes de registros eletrônicos de saúde, como os encontrados no banco de dados MIMIC-III [67].

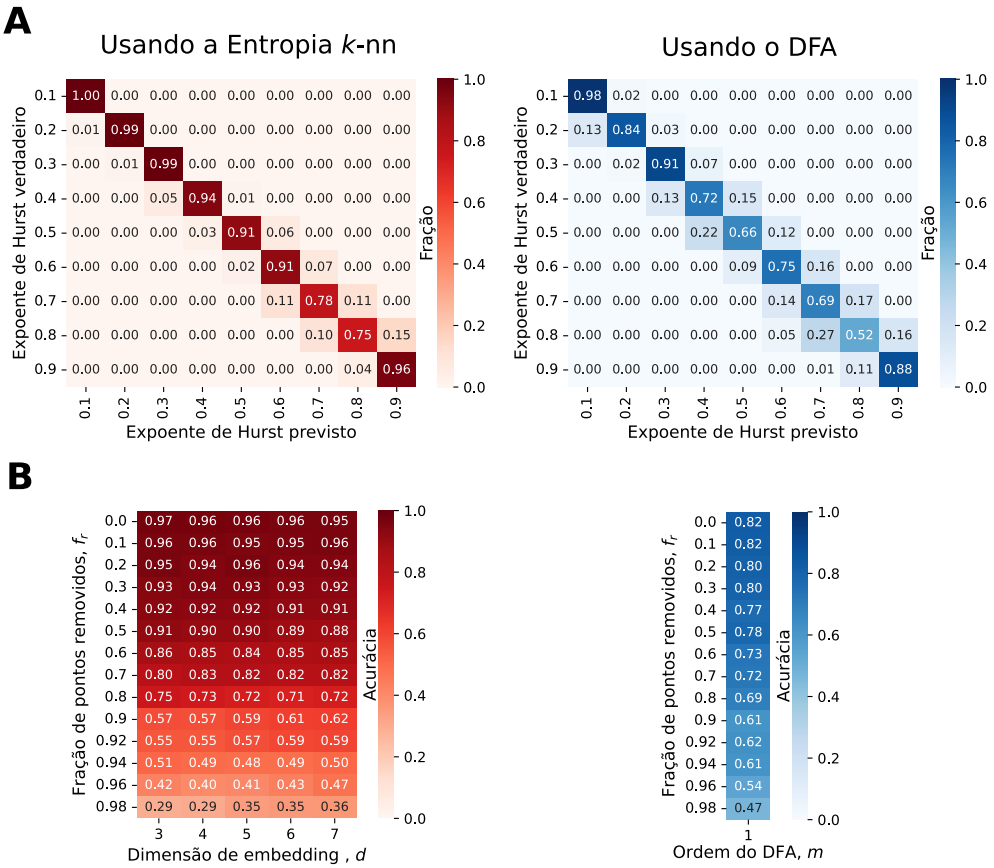

Figura 4.4: Detalhes das previsões e da acurácia das medidas de entropia de permutação k-nn e do DFA. (A) Exemplo de matrizes de confusão resultantes da aplicação do algoritmo de aprendizado para prever o valor de h usando a entropia S para $d = 3$ (tons de vermelho) e as estimativas do DFA h_{dfa} (tons de azul) com 45% dos pontos de dados excluídos aleatoriamente da série temporal. (B) Acurácia na previsão do expoente de Hurst h para várias frações de pontos de dados removidos f_r (linhas) usando a entropia S com diferentes dimensões *embedding* d (colunas em tons de vermelho) e usando o DFA (coluna em tons de azul). Nos dois painéis, as séries usadas tem tamanho $N = 2^{13} = 8192$.

4.2 Ruído harmônico

Em nossa segunda aplicação com séries temporais, examinamos dados geradas a partir de um processo estocástico chamado ruído harmônico [68], o qual corresponde a um oscilador harmônico com um ruído gaussiano aditivo. Esse processo é uma generalização do processo de Ornstein-Uhlenbeck [69] e pode ser expresso pelo sistema de equações de Langevin [68]

$$\begin{aligned} \frac{dx}{dt'} &= v \\ \frac{dv}{dt'} &= -\Gamma v - \Omega^2 x + \sqrt{2\varepsilon}\,\Omega^2 \xi(t') \end{aligned} , \tag{4.2}$$

no qual Γ, Ω e ϵ são parâmetros do modelo, enquanto que $\xi(t')$ é um ruído branco gaussiano. O ruído harmônico tem função de autocorrelação $C(\tau) = \langle x(t')x(t'+\tau)\rangle$ que exibe um decaimento exponencial oscilatório [68]

$$C(\tau) = \frac{\varepsilon\Omega^2}{\Gamma}\exp\left(-\frac{\Gamma}{2}\tau\right)\left[\cos(\omega\tau) + \frac{\Gamma}{2\omega}\sin(\omega\tau)\right], \tag{4.3}$$

com $\omega = \sqrt{\Omega^2 - (\Gamma/2)^2}$ sendo a frequência de oscilação. Além disso, no limite em que Γ e Ω tendem ao infinito mantendo a razão Γ/Ω^2 constante, esse ruído recai no processo de Ornstein-Uhlenbeck. Como pode ser visto na Figura 4.5, um ruído harmônico possui tanto comportamentos periódicos quanto comportamentos estocásticos a depender dos parâmetros do modelo. Esse comportamento também fica claro ao analisarmos a Equação 4.3, a qual descreve uma correlação que decai exponencialmente e oscila conforme decai, ou seja, os padrões estruturais das séries variam de acordo com ω. Fixando Γ e ϵ, a função oscila com pequenas variações para valores baixos de ω, caracterizando um regime semi-regular. No entanto, à medida que ω aumenta, a série se torna progressivamente mais aleatória. Logo, este é um bom modelo para verificar a capacidade de caracterização da entropia de permutação k-nn em séries temporais.

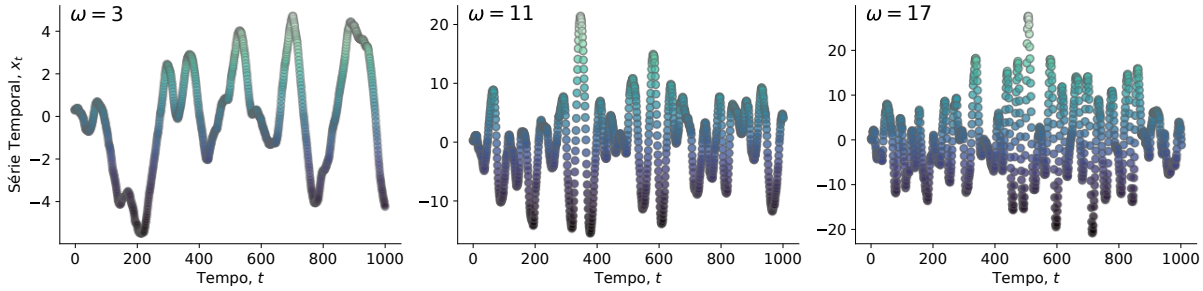

Figura 4.5: Sereis de ruído harmônico. Ilustrações de séries temporais de tamanho $N = 1000$ de ruídos harmônicos geradas com $\omega \in \{3, 11, 17\}$, $\Gamma = 3$ e $\epsilon = 1$. Para o valor mais baixo de ω, a série oscila de maneira relativamente regular; mas, à medida que ω aumenta, a série se torna mais aleatória.

Para gerar séries temporais do ruído harmônico, integramos numericamente as equações de Langevin 4.3 usando o método de Euler com um tamanho de passo $dt' = 10^{-2}$ para garantir que a função de autocorrelação exata e suas estimativas numéricas sejam mantidas. A integração prossegue até um tempo máximo de $t'_{\max} = 10$, resultando em séries temporais $\{x_t\}_{t=1,\ldots,N}$ com $N = 1000$ elementos. Geramos um conjunto composto por cem réplicas independentes dessas séries temporais fixando $\epsilon = 1$ e $\Gamma = 3$ para cada valor de ω variando entre 3 e 17 em incrementos de 2. Esses parâmetros produzem uma função de autocorrelação que se assemelha ao movimento de um oscilador harmônico subamortecido. Esse valores foram escolhidos porque a entropia de permutação usual apresenta dificuldades em identificar alterações na frequência de oscilação dentro dessa faixa [17], logo é um dado em que se pode

verificar se existe alguma diferença entre a entropia de permutação usual e a entropia de permutação k-nn.

Além disso, para avaliar a robustez da entropia de permutação dos k-primeiros vizinhos quanto à interferência por ruído externo e para poder compará-la com outras medidas, permutamos aleatoriamente pares de elementos nessas séries temporais em várias frações f_s do comprimento total, variando de 0 a 0.48 em 14 amostras com espaçamento crescente. A Figura 4.6 mostra exemplos dessas séries temporais para $\omega = 5$ com três diferentes frações f_s. Esses exemplos foram feitos a partir da mesma série de modo que a alguns aspectos da série original ainda são perceptíveis após embaralhar 50% de seus elementos.

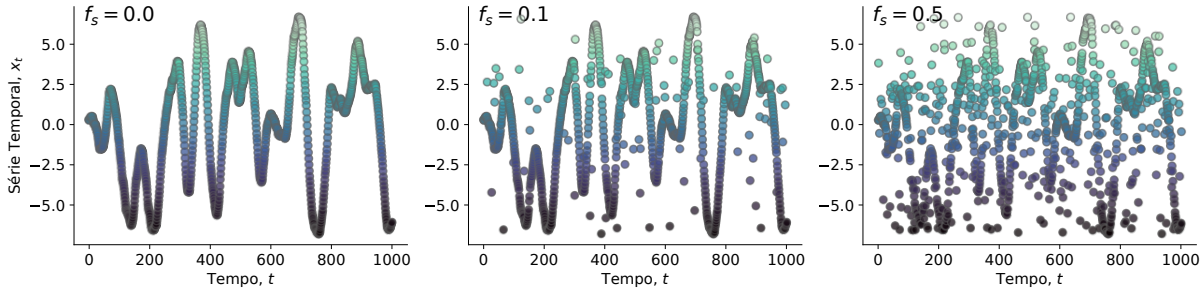

Figura 4.6: Sereis de ruído harmônico embaralhadas. Ilustrações de séries temporais de tamanho $N = 1000$ geradas a partir de ruídos harmônicos após trocar as posições de pares de pontos selecionados aleatoriamente em diferentes frações $f_s \in 0, 0.1, 0.5$. Os parâmetros utilizados foram $\omega = 5$, $\Gamma = 3$ e $\epsilon = 1$.

Usando esse processo, calculamos a dependência da entropia de permutação dos k-primeiros vizinhos com a frequência de oscilação ω para cada fração f_s, comparando os resultados com as relações derivadas da função de autocorrelação C_τ (estimada numericamente a partir de séries temporais) para diferentes atrasos τ e da entropia de permutação usual. A relação entre as medidas e o parâmetro ω, ilustradas na Figura 4.7A para $f_s = 0.01$ e $d = \tau = 3$, demonstram que tanto a entropia de permutação dos k-primeiros vizinhos quanto a entropia de permutação usual apresentam uma tendência crescente com os valores de ω, enquanto a autocorrelação diminui à medida que ω aumenta. Esse resultado concorda com o esperado qualitativo, uma vez que as entropias tendem a aumentar e a correlação a diminuir à medida que a estrutura na série temporal se torna mais irregular. No entanto, a taxa de variação relativa na entropia de permutação usual é consideravelmente menor do que aquelas observadas para as outras duas métricas. Os valores da entropia de permutação usual como função de ω estão quase inteiramente dentro dos intervalos de confiança de um desvio padrão, ressaltando a dificuldade de diferenciar ruídos harmônicos com essa quantidade [17]. Ao comparar as relações derivadas da entropia de permutação dos k-primeiros vizinhos e da autocorrelação, observamos que a dispersão relativa nos valores de S é substancialmente menor do que nos valores de C_τ, especialmente para valores de frequência mais altos.

Para comparar metodicamente as três medidas, aplicamos novamente nossa abordagem

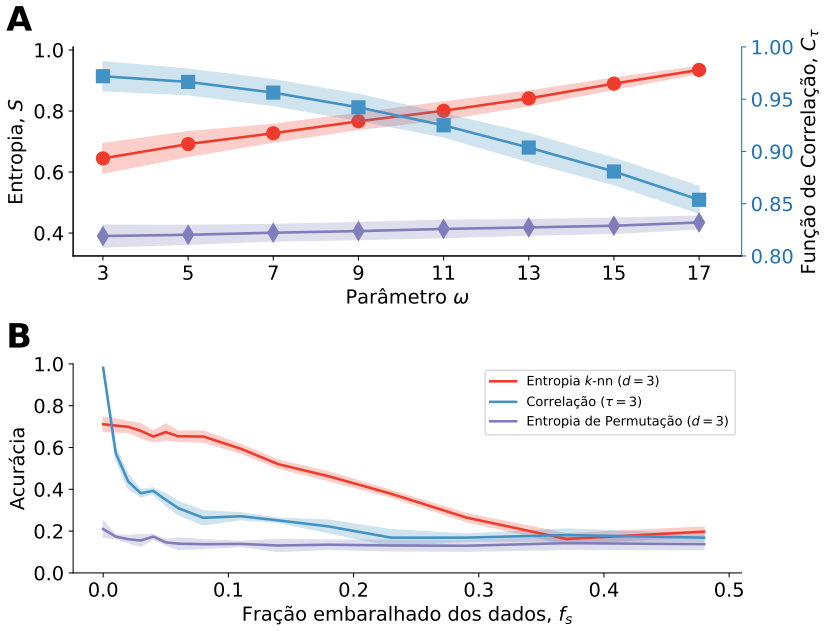

Figura 4.7: Comparação entre a entropia de permutação k-nn, a entropia de permutação usual e a função de correlação. (A) Relação entre a entropia de permutação dos k-primeiros vizinhos (S com $d = 3$) e a frequência do ruído harmônico ω (círculos vermelhos), justaposta à dependência da função de correlação C_τ (com atraso $\tau = 3$, quadrados azuis) e da entropia de permutação usual (com $d = 3$, losangos roxos) em relação ao parâmetro ω (ambos para $f_s = 0.01$). (B) Análise comparativa da acurácia na predição do parâmetro de frequência ω usando a entropia de permutação dos k-primeiros vizinhos (curva vermelha, $d = 3$), a função de correlação C_τ (curva azul, $\tau = 3$) e a entropia de permutação usual (curva roxa, $d = 3$), em função da fração de dados embaralhados f_s. As regiões sombreadas representam um desvio padrão nas estimativas de acurácia calculadas a partir de dez realizações independentes do processo de treinamento.

de aprendizado de máquina, usando cada uma das medidas para classificar os valores de ω em diferentes frações de dados embaralhados f_s. A Figura 4.7B mostra a acurácia média de cada classificador como uma função de f_s para $d = \tau = 3$. A entropia de permutação usual apresenta o pior desempenho, com níveis de acurácia consistentemente abaixo de $0,2$, um resultado que não melhora com maiores dimensões *embedding*. Sem ruído externo nas séries temporais ($f_s = 0$), os classificadores treinados com a função de autocorrelação C_τ atingem uma acurácia quase perfeita (97%) e superam significativamente aqueles treinados com a entropia de permutação dos k-primeiros vizinhos (69%). No entanto, a acurácia dos classificadores baseados no valor de C_τ é altamente suscetível ao ruído externo, caindo abaixo da entropia de permutação dos k-primeiros vizinhos quando apenas 1% dos pontos da série temporal são embaralhados. A acurácia dos classificadores baseados no valor de C_τ continua a diminuir para níveis comparáveis aos da entropia de permutação usual quando $f_s = 0,1$, enquanto a entropia de permutação dos k-primeiros vizinhos permanece robusta apesar da

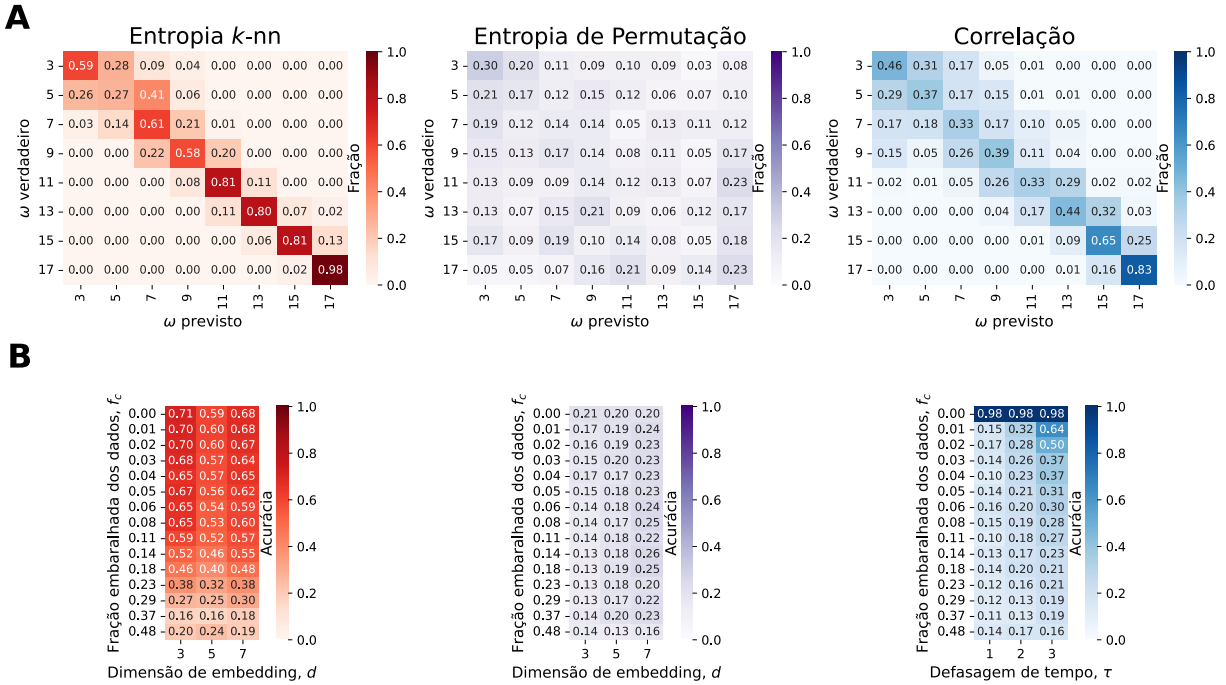

Figura 4.8: Detalhes das previsões e da acurácia das medidas de entropia de permutação k-nn, entropia de permutação usual e a função de correlação. (A) Exemplos de matrizes de confusão resultantes da aplicação do algoritmo de aprendizado para prever o parâmetro ω usando entropia de permutação dos k-primeiros vizinhos com $d = 3$ (tons de vermelho), entropia de permutação usual (tons de roxo) e a função de correlação com $\tau = 3$ (tons de azul) em séries temporais com 1% de dados embaralhados. (B) Avaliação da acurácia na previsão da frequência do ruído harmônico ($\omega \in 3, 5, \ldots, 17$) para diferentes valores de f_s (linhas) usando a entropia de permutação dos k-primeiros vizinhos com diferentes dimensões *embedding* (colunas, tons de vermelho), a entropia de permutação usual também com diferentes dimensões *embedding* (colunas, tons de roxo) e a função de correlação com diferentes intervalos de tempo (colunas, tons de azul).

adição de ruído externo. No limite em que quase um terço dos dados é alterado, nenhuma das medidas atinge uma acurácia significativa, o que é compreensível, já que nesse nível de embaralhamento os dados começam a perder qualquer traço das características gerais da série original.

Com ruído externo mínimo, $f_s = 0.01$, as matrizes de confusão dos classificadores treinados com as entropias e com a correlação são apresentadas na Figura 4.8A. Essas matrizes indicam que a entropia de permutação dos k-primeiros vizinhos melhora significativamente as classificações em valores de frequência mais baixos. A correlação parece se confundir nessa mesma região de valores baixos. A entropia de permutação confunde os valores quase que completamente. Como pode ser visto na Figura 4.8B, esses resultados não são significativamente influenciados por variações na dimensão de *embedding* d. O intervalo de tempo $\tau = 3$ parece resultar em uma melhor previsão em comparação aos outros para o caso da função de correlação.

A maior resistência a ruído da entropia de permutação dos k-primeiros vizinhos pode ser atribuível ao fato de que os padrões ordinais não são estritamente derivados de elementos sequenciais nas séries temporais. Em vez disso, eles são calculados a partir de amostras de séries temporais obtidas por meio de caminhadas aleatórias em um grafo de k-primeiros vizinhos que conecta pontos próximos dentro do espaço da série temporal, capturando, assim, os padrões dominantes da série temporal, mesmo em altos níveis de ruído externo. Pontos que se desviam significativamente desses padrões dominantes são amostrados com menor frequência pela trajetória do caminhante, aumentando a robustez ao ruído de nossa abordagem. A entropia de permutação, por outro lado, segue rigidamente a ordem dos pontos, e, portanto, não apresenta essa mesma robustez diante do embaralhamento dos dados

Dada essa robustez da entropia de permutação k-nn em lidar com ruído externo, uma aplicação promissora seria sua utilização na análise de séries temporais derivadas de dados empíricos que possuam componentes de ruído. Isso poderia ser especialmente útil em séries climáticas, como o índice de oscilação do Atlântico Norte (NAO) [70] e o índice de oscilação Sul (SOI) [71]. Esses índices representam oscilações climáticas de larga escala e são afetados por diversos fatores atmosféricos. Ao aplicar a entropia de permutação k-nn a esses dados, poderíamos identificar padrões e tendências subjacentes em fenômenos climáticos complexos.

Nesse contexto, a principal conclusão, e talvez a mais significativa, é que a entropia de permutação k-nn supera a entropia de permutação usual. Sendo assim, uma outra perspectiva é usá-la em conjunto com outras medidas estatísticas de complexidade para analisar séries provenientes de processos estocásticos ou caóticos, revisitando a caracterização com plano de complexidade-entropia usual [16, 17].

Conclusão

Nessa dissertação, introduzimos uma abordagem que generaliza o método de entropia de permutação para dados não estruturados. Nosso método baseia-se na construção de grafos de vizinhos mais próximos que conectam pontos de dados mais próximos e estabelecem relações de vizinhança entre eles. Caminhadas aleatórias sobre esses grafos amostram os valores associados a cada ponto de dados e permitem a extração de padrões ordinais e suas distribuições. A entropia de Shannon dessas distribuições ordinais define um quantificador que designamos como entropia de permutação dos k-vizinhos mais próximos ou entropia de permutação k-nn. Esta nova ferramenta permite a caracterização de tipos de dados além dos dados estruturados tradicionais (séries temporais ou imagens) que são tipicamente analisados com a entropia de permutação usual. Testamos a eficácia desta nova ferramenta ao examinar padrões em dados espaçados irregularmente derivados de experimentos *in silico* controlados, demonstrando sua eficácia em detectar mudanças nesses padrões e seu desempenho superior de classificação em comparação com uma medida amplamente utilizada de autocorrelação espacial conhecida como índice I de Moran.

Além de expandir a aplicabilidade dos métodos ordinais para dados não estruturados, também demonstramos que a entropia de permutação k-nn pode ser aplicada com sucesso em séries temporais e imagens. De fato, verificamos que os grafos de vizinhos mais próximos usados para calcular a entropia de permutação k-nn integram inerentemente informações sobre amplitude e intervalos de tempo na análise de dados regulares. Essa inclusão aumenta significativamente a resiliência ao ruído e a capacidade preditiva do nosso método em comparação com a entropia de permutação usual. Além disso, verificamos que a entropia de permutação k-nn supera métodos renomados em sistemas complexos, como a análise de flutuação destendenciada (DFA).

Por fim, acreditamos que a entropia de permutação k-nn tem potencial para a análise de tipos de dados multidimensionais, incluindo dados de nuvem de pontos, que são predo-

minantes em aplicações de sensores LiDAR e imagens médicas, bem como na caracterização de processos pontuais e dados geoespaciais. Nossa abordagem também abre caminho para a extensão de uma variedade de outros métodos ordinais para dados não estruturados, contribuindo com a necessidade premente de metodologias capazes de extrair informações interpretáveis e valiosas de conjuntos de dados cada vez mais volumosos e complexos usados na academia e na indústria.

APÊNDICE A

Aprendizado de máquina

De modo geral, o principal argumento de nosso trabalho é que a entropia de permutação dos k-primeiros vizinhos é capaz de distinguir entre dados regulares e irregulares, principalmente no caso de dados não estruturados. Esse argumento foi comprovado usando uma verificação por aprendizagem estatística, isto é, testamos se um modelo treinado com os valores da entropia de permutação k-nn poderia prever propriedades relacionadas à regularidade de dados simulados e reais.

Assim, suponha que tenhamos um *ensemble* com n amostras oriundas de algum sistema ou processo em que a irregularidade/regularidade de cada amostra é caracterizada por um parâmetro y, cujo conjunto, $Y = (y_1, y_2, \ldots y_n)$, representa as variáveis dependentes. Calculamos a entropia de cada amostra e agrupamos esses valores em um conjunto $X = (x_1, x_2, \ldots x_n)$ que representa as variáveis independentes. Desse modo, um modelo de aprendizado de máquina procura estimar uma relação $Y = f(X)$ entre as variáveis X e Y. Ao compararmos a previsão $f(X)$ feita pelo modelo com os valores de Y podemos verificar se realmente existe alguma relação [53]. No nosso caso em particular, queremos verificar se existe uma relação entre a entropia e os parâmetros que controlam a irregularidade do sistema.

Se as variáveis independentes assumirem valores numéricos contínuos, então o problema é uma tarefa de regressão. Por outro lado, se as variáveis independentes puderem ser agrupadas em classes ou em categorias, então o problema é uma tarefa de classificação. Além disso, a tarefa de previsão ainda pode ser supervisionada ou não supervisionada, sendo que na primeira a relação é estimada usando as duas variáveis, enquanto que na segunda apenas as variáveis independentes são usadas.

A.1 Classificação com os k-primeiros vizinhos

O modelo dos k-primeiros vizinhos (KNN) determina a classe de uma observação com base nos pontos de sua vizinhança [53]. Na fase de classificação, k é um parâmetro do modelo e um ponto de teste é classificado atribuindo a etiqueta que é mais frequente entre as k amostras de treinamento mais próximas daquele ponto.

Como pode ser visto na Figura A.1, a previsão funciona a partir de uma votação dos vizinhos mais próximos (ilustrados com $k = 2$ e $k = 5$). Nesse exemplo, temos duas classes: os círculos em verde e os círculos em roxo. O modelo vai prever a qual classe pertence o círculo azul (que representa um dado não rotulado). A classificação usa os vizinhos mais próximos, na qual uma "votação" é realizada entre esses vizinhos e a classe mais representada é atribuída ao dado sem rótulo. No caso de $k = 2$, existe um representante de cada classe, resultando em um empate que é resolvido de maneira aleatória. Quando $k = 5$, temos três representantes da classe verde e dois da roxa entre os vizinhos do dado sem rótulo, logo a votação fica a favor da verde e essa será a cor que o preditor vai associar ao círculo azul.

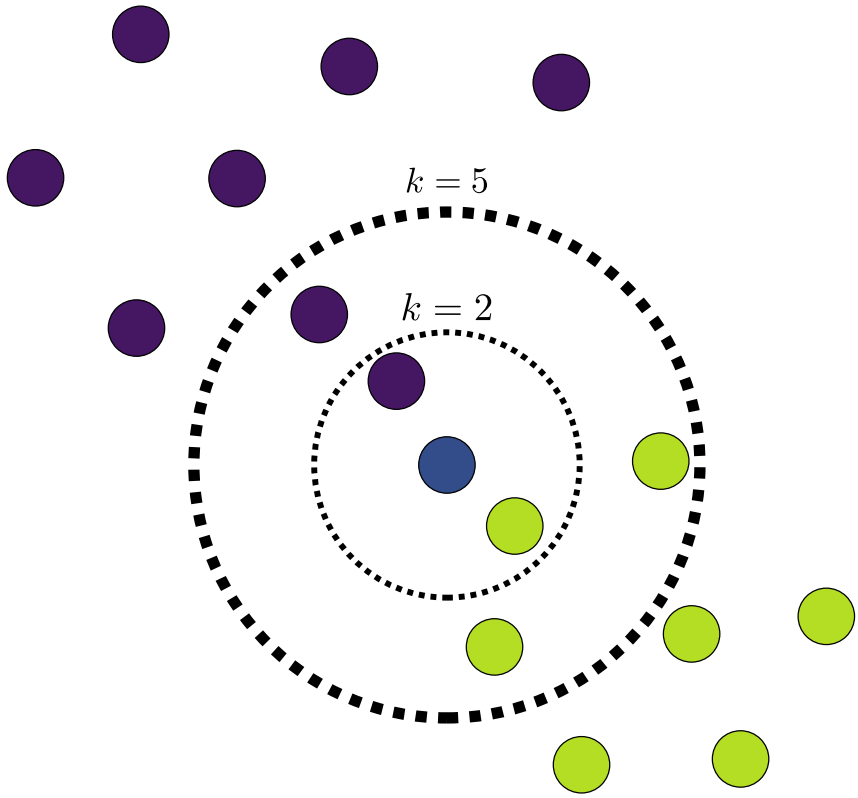

Figura A.1: Modelo de classificação dos k-primeiros vizinhos (KNN). Exemplo de classificação baseado no modelo KNN. Em azul está o ponto cuja categoria se deseja prever. Os pontos verdes e os pontos roxos representam duas categorias diferentes. Os círculos pontilhados indicam a região que engloba o número k de vizinhos escolhido (aqui ilustrado com $k = 2$ e $k = 3$. Para dois vizinhos, temos um representante de cada classe. Para cinco vizinhos, temos três representantes de uma classe e dois da outra, logo o modelo vai atribuir ao ponto não classificado a classe verde.

De maneira mais formal, o modelo KNN identifica os k vizinhos mais próximos baseado na distância euclidiana de um ponto x_0 e estima a probabilidade condicional do ponto pertencer à classe j, isto é,

$$P(Y = j|X = x_0) = \frac{1}{k} \sum_{x_i} \in \mathbf{N_0} I y_i = j \tag{A.1}$$

na qual $\mathbf{N_0}$ é o conjunto de vizinhos e $Iy_i = j$ é uma função indicadora que vale 1 se o ponto x_i pertence à classe j e 0 caso contrário. A previsão do modelo corresponde à classe com maior probabilidade.

A escolha do valor de k é crucial para o desempenho do modelo KNN. Valores pequenos de k podem ser muito sensíveis ao ruído nos dados, enquanto valores grandes podem suavizar demais a decisão, levando a classificações incorretas. Embora a distância euclidiana seja a mais comum, outros tipos de distâncias, como a distância de Manhattan, a distância de Minkowski, e a distância de Mahalanobis, também podem ser utilizadas dependendo das características dos dados. É importante normalizar ou padronizar os dados antes de aplicar o KNN, especialmente se as características tiverem escalas diferentes. Caso contrário, as características com maiores magnitudes podem dominar a medida de distância.

O KNN pode ser computacionalmente intensivo, especialmente para grandes conjuntos de dados, pois envolve calcular a distância entre o ponto de teste e todos os pontos de treinamento. No entanto, é amplamente utilizado em reconhecimento de padrões, recuperação de informações, detecção de anomalias e em sistemas de recomendações devido à sua simplicidade e eficácia.

A.2 O modelo *XGBoost*

O *XGBoost* (*extreme gradient boosting*) é um algoritmo de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão que é conhecido por sua eficiência, flexibilidade e desempenho superior em competições de ciência de dados. Sua implementação, que otimizada do algoritmo de *gradient boosting* combina o poder de múltiplos "modelos fracos" para formar um "modelo forte". Utilizando uma abordagem de *boosting*, o XGBoost treina vários modelos fracos (tipicamente árvores de decisão) sequencialmente. Cada novo modelo tenta corrigir os erros cometidos pelos modelos anteriores, ajustando-se para os resíduos das previsões anteriores. O processo iterativo continua até que o erro total seja minimizado ou um número predefinido de iterações seja alcançado.

Primeiramente, iniciamos com um modelo inicial, geralmente uma previsão constante. Em seguida, calculamos os resíduos que são as diferenças entre as previsões atuais e os valores reais. Depois disso, novos modelos fracos são treinados para prever esses resíduos. O modelo principal é então atualizado combinando os novos modelos fracos com pesos apropriados. Esse processo é repetido até que o critério de parada seja atingido. Essa abordagem iterativa

permite que o *XGBoost* melhore continuamente suas previsões ao longo do tempo.

O *XGBoost* oferece várias vantagens que o tornam uma escolha popular para tarefas de aprendizado de máquina. Primeiramente, ele é altamente otimizado do ponto de vista computacional, aproveitando técnicas como paralelização e manipulação eficiente de dados esparsos. Além disso, o *XGBoost* inclui termos de regularização que penalizam a complexidade dos modelos, ajudando a prevenir o *overfitting* aos dados. O algoritmo também possui mecanismos internos para lidar com valores faltantes, tornando-o robusto em cenários nos quais os dados podem ser incompletos. Sua flexibilidade é outra vantagem, já que suporta uma ampla gama de funções de perda e pode ser usado para classificação e regressão.

A.3 Acurácia e matriz de confusão

Uma das maneiras de quantificar a acurácia das previsões de modelos de aprendizado de máquina e utilizar a métrica de acurácia. Essa medida é definida como a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas. Em outras palavras, a acurácia mede a capacidade do modelo de classificar corretamente as amostras em suas respectivas classes. Essa métrica é especialmente útil quando as classes estão equilibradas, ou seja, quando o número de amostras em cada classe é aproximadamente o mesmo (como ocorreu em todas as nossas aplicações). Considerando que temos duas classes denominadas positivas e negativas, podemos definir a acurácia como

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.2}$$

na qual TP são as previsões corretas de positivas, TN são as previsões corretas de negativas, FP são previsões incorretas de positivas e FN as previsões incorretas de negativas.

A matriz de confusão é outra ferramenta útil na avaliação do desempenho de modelos de classificação em aprendizado de máquina. Ela fornece uma representação visual das previsões feitas pelo modelo, comparando-as com os valores reais. Essa matriz permite uma análise detalhada de como o modelo está se comportando em termos de previsões corretas e incorretas, fornecendo informações sobre onde ele está acertando e onde está cometendo erros. A matriz de confusão é particularmente útil em situações nas quais há um desbalanceamento nas classes, ou seja, quando o número de amostras em cada classe não é aproximadamente o mesmo. Nessas situações, a acurácia pode ser enganadora e a matriz de confusão permite uma compreensão mais clara de como o modelo está se comportando em relação a cada classe. A Figura A.2 mostra uma matriz de confusão de uma tarefa de classificação binária, na qual podemos identificar padrões de erro específicos que podem ser corrigidos para melhorar o desempenho do modelo.

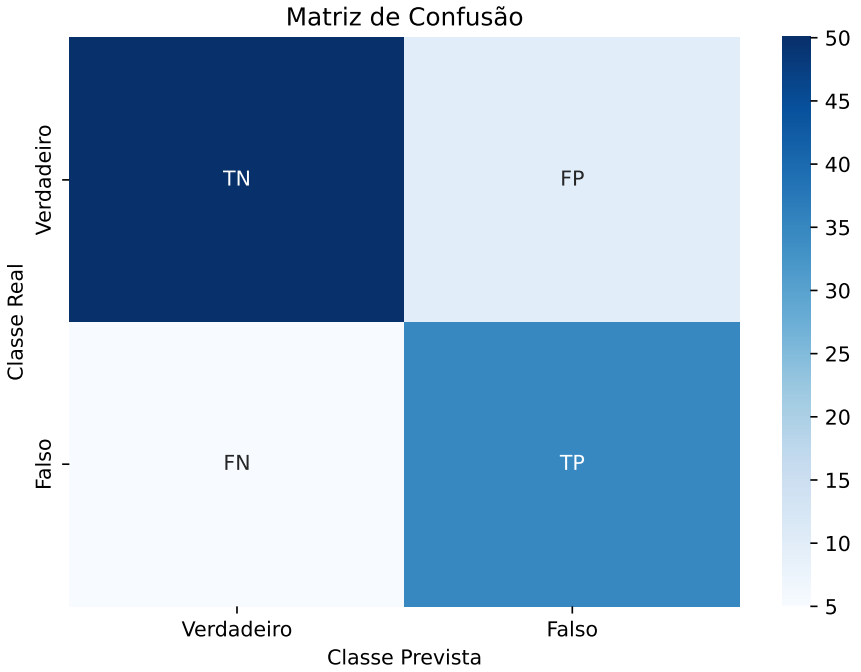

Figura A.2: Matriz de confusão de uma classificação binário. Exemplo de matriz de confusão para classificações binárias, na qual os tons de azul representam a proporção hipotética de ocorrência de cada elemento da matriz.

Referências Bibliográficas

[1] Mattmann, C. A. A vision for data science. *Nature* **493**, 473–475 (2013).

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

[3] Yan, R., Liu, Y. & Gao, R. X. Permutation entropy: A nonlinear statistical measure for status characterization of rotary machines. *Mechanical Systems and Signal Processing* **29**, 474–484 (2012).

[4] Nicolaou, N. & Georgiou, J. Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. *Expert Systems with Applications* **39**, 202–209 (2012).

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

[6] Garland, J., Jones, T. R., Neuder, M., Morris, V., White, J. W. & Bradley, E. Anomaly detection in paleoclimate records using permutation entropy. *Entropy* **20**, 931 (2018).

[7] Sigaki, H. Y., De Souza, R., De Souza, R., Zola, R. & Ribeiro, H. Estimating physical properties from liquid crystal textures via machine learning and complexity-entropy methods. *Physical Review E* **99**, 013311 (2019).

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

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

[10] Zanin, M., Zunino, L., Rosso, O. A. & Papo, D. Permutation entropy and its main biomedical and econophysics applications: A review. *Entropy* **14**, 1553–1577 (2012).

[11] Riedl, M., Müller, A. & Wessel, N. Practical considerations of permutation entropy. *The European Physical Journal Special Topics* **222**, 249–262 (2013).

[12] Amigó, J. M., Keller, K. & Unakafova, V. A. Ordinal symbolic analysis and its application to biomedical recordings. *Philosophical Transactions of the Royal Society A* **373**, 20140091 (2015).

[13] Keller, K., Mangold, T., Stolz, I. & Werner, J. Permutation entropy: New ideas and challenges. *Entropy* **19**, 134 (2017).

[14] Pessa, A. A. B. & Ribeiro, H. V. ordpy: A python package for data analysis with permutation entropy and ordinal network methods. *Chaos* **31**, 063110 (2021).

[15] Unakafov, A. M. & Keller, K. Conditional entropy of ordinal patterns. *Physica D* **269**, 94–102 (2014).

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

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

[18] Amigó, J. M., Kocarev, L. & Szczepanski, J. Order patterns and chaos. *Physics Letters A* **355**, 27–31 (2006).

[19] Amigó, J. M., Zambrano, S. & Sanjuán, M. A. True and false forbidden patterns in deterministic and random dynamics. *EPL* **79**, 50001 (2007).

[20] Bian, C., Qin, C., Ma, Q. D. Y. & Shen, Q. Modified permutation-entropy analysis of heartbeat dynamics. *Physical Review E* **85**, 021906 (2012).

[21] Cuesta-Frau, D., Varela-Entrecanales, M., Molina-Picó, A. & Vargas, B. Patterns with equal values in permutation entropy: Do they really matter for biosignal classification? *Complexity* **2018**, 1324696 (2018).

[22] Zunino, L., Soriano, M. C., Fischer, I., Rosso, O. A. & Mirasso, C. R. Permutation-information-theory approach to unveil delay dynamics from time-series analysis. *Physical Review E* **82**, 046212 (2010).

[23] Small, M. Complex networks from time series: Capturing dynamics. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, 2509–2512 (2013).

[24] McCullough, M., Small, M., Stemler, T. & Iu, H. H.-C. Time lagged ordinal partition networks for capturing dynamics of continuous dynamical systems. *Chaos* **25**, 053101 (2015).

[25] Small, M., McCullough, M. & Sakellariou, K. Ordinal network measures — quantifying determinism in data. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5 (2018).

[26] Pessa, A. A. B. & Ribeiro, H. V. Characterizing stochastic time series with ordinal networks. *Physical Review E* **100**, 042304 (2019).

[27] Borges, J. B., Ramos, H. S., Mini, R. A. F., Rosso, O. A., Frery, A. C. & Loureiro, A. A. F. Learning and distinguishing time series dynamics via ordinal patterns transition graphs. *Applied Mathematics and Computation* **362**, 124554 (2019).

[28] Pessa, A. A. B. & Ribeiro, H. V. Mapping images into ordinal networks. *Physical Review E* **102**, 052312 (2020).

[29] 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).

[30] Zunino, L. & Ribeiro, H. V. Discriminating image textures with the multiscale two-dimensional complexity-entropy causality plane. *Chaos, Solitons & Fractals* **91**, 679–688 (2016).

[31] Bandt, C. & Wittfeld, K. Two new parameters for the ordinal analysis of images. *Chaos* **33**, 043124 (2023).

[32] Voltarelli, L. G., Pessa, A. A., Zunino, L., Zola, R. S., Lenzi, E. K., Perc, M. & Ribeiro, H. V. Characterizing unstructured data with the nearest neighbor permutation entropy. *Chaos* **34** (2024).

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

[34] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* 1802.03426 (2018).

[35] McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3**, 861 (2018).

[36] Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 855–864 (2016).

[37] Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

[38] Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C. & Li, L. Rolx: structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1231–1239 (2012).

[39] Cox, D. R. & Isham, V. *Point processes*, vol. 12 (CRC Press, 1980).

[40] Rosso, O. A., Ospina, R. & Frery, A. C. Classification and verification of handwritten signatures with time causal information theory quantifiers. *PLOS ONE* **11**, e0166868 (2016).

[41] Mandelbrot, B. B. *The Fractal Geometry of Nature* (Freeman, New York, San Francisco, 1982).

[42] Mandelbrot, B. B. & Van Ness, J. W. Fractional Brownian motions, fractional noises and applications. *SIAM Review* **10**, 422–437 (1968).

[43] Hosking, J. R. M. Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research* **20**, 1898–1908 (1984).

[44] Burnecki, K., Kepten, E., Janczura, J., Bronshtein, I., Garini, Y. & Weron, A. Universal algorithm for identification of fractional Brownian motion. A case of telomere subdiffusion. *Biophysical Journal* **103**, 1839–1847 (2012).

[45] Stadler, L. & Weiss, M. Non-equilibrium forces drive the anomalous diffusion of telomeres in the nucleus of mammalian cells. *New Journal of Physics* **19**, 113048 (2017).

[46] Weber, S. C., Spakowitz, A. J. & Theriot, J. A. Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. *Physical Review Letters.* **104**, 238102 (2010).

[47] Bronshtein, I. *et al.* Loss of lamin a function increases chromatin dynamics in the nuclear interior. *Nature Communications* **6**, 8044 (2015).

[48] Jeon, J.-H., Tejedor, V., Burov, S., Barkai, E., Selhuber-Unkel, C., Berg-Sørensen, K., Oddershede, L. & Metzler, R. In vivo anomalous diffusion and weak ergodicity breaking of lipid granules. *Physical Review Letters* **106**, 048103 (2011).

[49] Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).

[50] Getis, A. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis* **40**, 297–309 (2008).

[51] de Jong, P., Sprenger, C. & van Veen, F. On extreme values of Moran's I and Geary's c. *Geographical Analysis* **16**, 17–24 (1984).

[52] Gittleman, J. L. & Kot, M. Adaptation: Statistics and a null model for estimating phylogenetic effects. *Systematic Zoology* **39**, 227–241 (1990).

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

[54] Yeung, D.-Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T. & Rigoll, G. Svc2004: First international signature verification competition. In *Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004. Proceedings*, 16–22 (Springer, 2004).

[55] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).

[56] Van der Gon Denier, J. & Thuring, J. P. The guiding of human writing movements. *Kybernetik* **2**, 145–148 (1965).

[57] Nalwa, V. S. Automatic on-line signature verification. *Proceedings of the IEEE* **85**, 215–239 (1997).

[58] De Gennes, P.-G. & Prost, J. *The Physics of Liquid Crystals*. 83 (Oxford University Press, 1993).

[59] Osipov, M. & Pikin, S. Dipolar and quadrupolar ordering in ferroelectric liquid crystals. *Journal de Physique II* **5**, 1223–1240 (1995).

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

[61] Yang, D.-K. & Wu, S.-T. *Fundamentals of Liquid Crystal Devices* (John Wiley & Sons, 2014).

[62] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. & Goldberger, A. L. Mosaic organization of DNA nucleotides. *Physical Review E* **49**, 1685–1689 (1994).

[63] Shao, Y.-H., Gu, G.-F., Jiang, Z.-Q., Zhou, W.-X. & Sornette, D. Comparing the performance of FA, DFA and DMA using different synthetic long-range correlated time series. *Scientific Reports* **2**, 835 (2012).

[64] Schulz, M. & Stattegger, K. Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series. *Computers & Geosciences* **23**, 929–945 (1997).

[65] Foster, G. Wavelets for period analysis of unevenly sampled time series. *Astronomical Journal* **112**, 1709–1729 (1996).

[66] Rehfeld, K., Marwan, N., Heitzig, J. & Kurths, J. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics* **18**, 389–404 (2011).

[67] Johnson, A. E., Stone, D. J., Celi, L. A. & Pollard, T. J. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association* **25**, 32–39 (2018).

[68] L. Schimansky-Geier, C. Z. Harmonic noise: Effect on bistable systems. *Zeitschrift für Physik B* **79**, 451–460 (1990).

[69] Gardiner, C. W. *et al. Handbook of Stochastic Methods*, vol. 3 (Springer Berlin, 1985).

[70] Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. An overview of the north atlantic oscillation. *Geophysical Monograph-American Geophysical Union* **134**, 1–36 (2003).

[71] Trenberth, K. E. Signal versus noise in the southern oscillation. *Monthly Weather Review* **112**, 326–332 (1984).