VETOR

Mapas de difusão na busca de *outliers* de curvas de polarização☆

Diffusion maps in search of outliers of polarization curves

Carlos Alberto Lopes dos Santos de Oliveira[1,†], Ivan Napoleão Bastos[2], Francisco Duarte Moura Neto[2]

[1]*Instituto Federal do Norte de Minas Gerais, Campus Teófilo Otoni - Teófilo Otoni, MG, Brasil*
[2]*Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, Brasil*
[†]**Autor correspondente:** carlos.santos@ifnmg.edu.br

Resumo

Este trabalho propõe uma abordagem de uma eficiente técnica de redução de dimensionalidade não linear conhecida como mapas de difusão na identificação de *outliers* de curvas de polarização. Tais curvas avaliam a cinética global de processos eletroquímicos de eletrodos em meios corrosivos sob condições estacionárias, mas que apresentam variações de natureza experimental ou por erros sistemáticos que possam ocorrer. Assim, a busca por *outliers* se faz necessária pois, no contexto em questão, permitiria extrair do conjunto das curvas experimentais em análise aquelas que, eventualmente, não traduzem de forma adequada o comportamento do material imerso no meio corrosivo a um dado potencial aplicado. A esperança é conseguir um ganho na taxa de classificação, eliminando-se dados muito improváveis do conjunto de amostras. Classificá-los corretamente é um desafio, pois a forte não linearidade típica de curvas de polarização e a superposição tornam a tarefa difícil. Neste trabalho, estudou-se dois aços inoxidáveis em solução aquosa com 3, 5% NaCl. Por meio da matriz de difusão e do mapa de cores a ela associada, a técnica de mapas de difusão conseguiu encontrar satisfatoriamente *outliers* de ambos os aços utilizados, mostrando-se útil diante da abordagem proposta.

Palavras-chave

Mapas de difusão • Outliers • Curvas de polarização

Abstract

This work proposes an efficient nonlinear dimensionality reduction technique known as Diffusion Maps in the identification of polarization curve outliers. Such curves evaluate the global kinetics of electrochemical electrode processes in corrosive media under steady-state conditions. However, they can present scattering of experimental nature or due to artifacts that may occur. Thus, a search for outliers is necessary because it would allow extracting them from the set of experimental curves under analysis; eventually, they do not adequately translate the material's behavior immersed in the corrosive environment to a given electrochemical potential. The hope is to achieve a good classification rate, eliminating improbable results from the data. However, classifying them correctly is a challenge because a strong nonlinearity typical of polarization curves besides the overlap make the task difficult. In this work, two stainless steels were studied in an aqueous solution with 3.5% NaCl. Through the diffusion matrix and the associated color map, the Diffusion Maps technique was able to satisfactorily find outliers for both steels used, proving useful because of the proposed approach.

Keywords

Diffusion Maps • Outliers • Polarization curves

☆Este artigo é uma versão estendida do trabalho apresentado no XXIII ENMC Encontro Nacional de Modelagem Computacional e XI ECTM Encontro de Ciência e Tecnologia de Materiais, ocorrido em modo *webinar*, de 20 a 30 de outubro de 2020.

1 Introdução

Por meio da experimentação, informações quantitativas (ou qualitativas) são extraídas da natureza pelas relações existentes entre os fenômenos observados, expressando-as mediante grandezas físicas numéricas ou convertidas numericamente. Neste contexto, muitos dados oriundos de técnicas experimentais em ciência e engenharia de materiais estão disponíveis na forma de perfis ([1]), pois dependem fortemente de grandezas que variam sob uma faixa contínua, como a energia, comprimento de onda, temperatura, potencial dentre outros.

No presente trabalho, fazemos a análise de curvas de polarização de corrosão. Tratam-se, na verdade, de discretizações de curvas de polarização que passam a ser vistos como vetores. Essas curvas, por sua vez, de maneira mais geral, são parcelas com potencial de corrente e têm fundamental importância nos estudos de corrosão e de eletroquímica, sendo essenciais para medir a cinética global dos eletrodos.

Ao dispor desses dados experimentais, naturalmente é de se imaginar que esses são compostos de uma parcela de ruído, decorrente da experimentação, e que, eventualmente, algumas observações podem não traduzir de forma adequada o comportamento dos aços em estudo em relação ao seu potencial de corrosão localizada. A esses perfis, doravante, referimos como *outliers* da amostra.

A busca por *outliers* tem considerável importância. O estabelecimento de um perfil base de um produto ou serviço, por exemplo, em inúmeras situações, garante a melhoria da sua qualidade a medida que serve de modelo para o emprego de ferramentas de monitoramento estatístico nas indústrias de manufatura e serviços. Segundo Wang *et al.*[2], o monitoramento do processo é uma questão importante para garantir a segurança e operações eficientes nas indústrias de processo modernas.

Os dados utilizados neste trabalho são os mesmos utilizados em Fabbri *et al.* [1]: dois aços inoxidáveis austeníticos comerciais (UNS S30400 e UNS S31600) submetidos à polarização potenciodinâmica que resulta nas curvas de polarização. Neste artigo de 2019, basicamente, o objetivo foi avaliar a abordagem multi-q, usando a entropia de Tsallis, na classificação dos perfis com a separação eficiente.

Neste trabalho, por sua vez, é utilizada a técnica de mapas de difusão na busca por *outliers*, fato que, aparentemente, possibilitaria um ganho na classificação em relação a todo o conjunto de dados. Se pudéssemos identificar inicialmente os perfis *outliers*, poderíamos suprimi-los e reavaliar a classificação para o novo conjunto contendo apenas dados válidos. A esperança é que, mesmo ainda não realizando a classificação neste trabalho, haja uma depuração dos dados. A técnica foi utilizada com sucesso, por exemplo, na obtenção de perfis provenientes da produção de madeira processada em relação à densidade ao longo da espessura por Moura Neto *et al* [3].

Inicialmente, a seção 2 traz uma breve apresentação da técnica. O objetivo é possibilitar ao leitor compreender seu funcionamento e exibir os passos para sua implementação. Na seção 3, é aplicada a técnica às curvas de polarização na identificação dos *outliers*. Por fim, a última seção expõe as considerações finais e expectativas de estudos posteriores.

2 Modelagem Matemática

2.1 Mapas de difusão

Os mapas de difusão estão entre as mais recentes e produtivas técnicas para reduzir a dimensionalidade não linear de dados, além de permitir localizar as estruturas importantes. Introduzido por Coifman e Lafon [4] e por Coifman e Hirn [5], o método, que emprega algoritmos baseados em grafos, permite extrair informações relevantes sobre os dados em estudo proporcionando uma melhor compreensão sobre o fenômeno investigado. Ao utilizar grafos para representar conjuntos de dados e suas semelhanças, esta técnica permite definir uma métrica no espaço de distribuição de probabilidades que usa de processos markovianos para correlacionar similaridades de diferentes ordens e escalas a partir das coordenadas desses pontos e da noção de afinidade estabelecida. De acordo com Lafon [6], neste espaço os pontos de dados são reorganizados de tal forma que a geometria combina todas as informações locais captadas pelo processo de difusão, e a distância euclidiana define uma métrica de difusão que mede a proximidade dos pontos em termos de sua conectividade. O restante desta seção aborda em detalhes os aspectos do algoritmo.

2.2 Conectividade

O passo inicial na implementação do algoritmo é medir a conectividade entre os pontos de dados observados. Neste sentido, conectividade é sinônimo de similaridade. Imagine cada ponto como o vértice de um grafo conectado com arestas que o ligam a outro nó vizinho têm seus pesos definidos de acordo com a similaridade que cada um destes tem com seus vizinhos. Tem-se, então, um grafo com pesos que refletem uma similaridade local entre os pontos de dados.

Façamos agora um passeio aleatório pelos dados, começando de um específico, fazendo saltos aleatórios, guiado apenas pela probabilidade de atingir um ponto qualquer. Saltar para um ponto mais próximo é mais provável do que

saltar para outro que esteja distante. Feito desse modo, cria-se uma relação entre a distância no espaço de dados e a probabilidade de transição do caminhante aleatório.

A conectividade entre dois pontos de dados quaisquer, $\mathbf{x}_i$ e $\mathbf{x}_j$, com $i \neq j$ é definida como a probabilidade de saltar de $\mathbf{x}_i$ para $\mathbf{x}_j$ em cada instante da caminhada aleatória. Como essa conectividade é proporcional à similaridade dos dados, é útil expressar essa conectividade em termos de uma função de verossimilhança não normalizada, $\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$, conhecida como núcleo de difusão (*diffusion kernel*).

$$\text{conectividade}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) \tag{1}$$

Este núcleo define uma medida local de similaridade dentro de uma certa vizinhança. Fora dela, esta função precisa se aproximar rapidamente para zero. Uma função que atende muito bem a este propósito é o núcleo gaussiano Eq. (2).

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\alpha}\right) \tag{2}$$

A noção de vizinhança de um dado qualquer $\mathbf{x}_i$ pode ser definida como todos os pontos $\mathbf{x}_n$, com $n \in \mathbb{N}$, cuja a função $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) \geq \varepsilon$ escolhido. Esse parâmetro remete à conectividade do grafo e como o pesquisador imagina que os dados possam estar relacionados. Fazendo-se o ajuste da escala do núcleo (parâmetro α), escolhe-se o tamanho da vizinhança de acordo com o conhecimento a priori da disposição e da densidade dos dados.

Seja a matriz $\tilde{K}$ com os valores de conectividade entre todos os pontos de dados. Definimos agora uma distribuição de difusão normalizada por uma uma matriz P, conhecida como matriz de Markov (ou matriz de probabilidade de transição) cujas entradas representam a probabilidade de o caminhante aleatório atingir cada ponto de dado em determinado instante.

Obtém-se esta matriz P utilizando-se da matriz de similaridade $\tilde{K}$ como se segue:

$$(K)_{ij} = k_{ij} = \frac{\tilde{k}_{ij}}{\tilde{k}_i \tilde{k}_j}, \qquad \tilde{k}_i = \sum_{w=1}^{n} \tilde{k}_{iw} \tag{3}$$

$$(P)_{ij} = p_{ij} = \frac{k_{ij}}{m_i}, \qquad m_i = \sum_{w=1}^{n} k_{iw} \tag{4}$$

A matriz M é uma matriz diagonal que tem em cada linha, em sua entrada não-nula, a soma dos pesos das linhas correspondente da matriz de conectividade K ajustada. Se por um lado, P não é simétrica, em contrapartida ganha-se uma propriedade de conservação (a soma das linhas é 1) e representa, em cada entrada p_{ij}, a probabilidade de transição de um vértice $\mathbf{x}_i$ a outro $\mathbf{x}_j$ em uma unidade de tempo. É interessante ainda observar que a analogia entre o processo descrito e as cadeias de Markov permite identificar cada ponto de dado como um estado desta cadeia.

2.3 Processo de Difusão

Como observado, $p(\mathbf{x}_i, \mathbf{x}_j)$ representa a probabilidade de transição em um passo de tempo do estado $\mathbf{x}_i$ para o $\mathbf{x}_j$ e é proporcional ao peso da aresta $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$. Ao tomar as potências da matriz de difusão P, aumenta-se o número de passos percorridos e o processo de difusão prossegue. À medida que a cadeia avança no tempo, o que equivale a tomar cada vez maiores potências de P, permite-se integrar a geometria local e, com isso, ter conhecimento sobre as estruturas geométricas da variedade na qual o conjunto de dados está inserido.

Na seção seguinte, observa-se que é a matriz de difusão P a responsável pelas informações que possibilitam encontrar os *outliers* no conjunto. Ao ordenar as colunas e linhas da matriz, segundo o primeiro autovetor não constante, consegue-se uma maneira de agrupar os perfis, do ponto de vista das cadeias de Markov, obtendo subconjuntos de perfis cuja probabilidade de algum dado escapar é baixa, isso em diferentes instantes de tempo.

2.4 Distâncias e mapas de difusão

O processo de difusão utilizando as potências da matriz Markov P deve ser útil na descoberta da estrutura geométrica que contém os dados em várias escalas. Define-se uma métrica com base nesta estrutura. Seja esta a medida da similaridade de dois pontos de dados nesse espaço como a conectividade entre eles. Esta grandeza relaciona-se com a matriz P sendo dada por:

$$D_t(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{y \in X} \left|p_t(\mathbf{x}_j, y) - p_t(\mathbf{x}_i, y)\right|^2 = \sum_k \left|P^t_{jk} - P^t_{ik}\right|^2 \tag{5}$$

Esta distância será pequena se houver muitos caminhos de alta probabilidade com comprimento t entre dois pontos e permanecerá pequena enquanto as probabilidades de caminho entre $\mathbf{x}_i$ e $\mathbf{x}_j$ continuarem pequenas. Como a distância leva em conta a soma entre todos os caminhos entre os pontos, a métrica deve ser robusta à perturbação por ruídos.

É apresentado agora como mapear os pontos de dados para um espaço euclidiano de acordo com a métrica definida. A distância de difusão no espaço original torna-se a distância euclidiana neste novo espaço de difusão. Seja o mapeamento:

$$Z_i = \begin{bmatrix} p_t(\mathbf{x}_i, \mathbf{x}_1) \\ p_t(\mathbf{x}_i, \mathbf{x}_2) \\ \vdots \\ p_t(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \tag{6}$$

Definido deste modo, a distância euclidiana entre dois pontos mapeados, Z_i e Z_j, é:

$$\left\|Z_j - Z_i\right\|^2 = \sum_{y \in X} \left|p_t(\mathbf{x}_j, y) - p_t(\mathbf{x}_i, y)\right|^2 = \sum_k \left|P^t_{jk} - P^t_{ik}\right|^2 = D_t(\mathbf{x}_i, \mathbf{x}_j)^2 \tag{7}$$

que, como visto, é a distância de difusão entre os pontos de dados $\mathbf{x}_i$ e $\mathbf{x}_j$.

A redução é realizada negligenciando-se certas dimensões no espaço de difusão. Mas antes, é necessário apresentar uma alternativa para os cálculos da distância de difusão. Seja $\boldsymbol{\psi}_n$ os autovetores direitos de P associados com os autovalores $\lambda_k, k = 1, \ldots, n$. A distância de difusão entre $\mathbf{x}_i$ e $\mathbf{x}_j$ para um t fixo pode ser calculada como:

$$D_t(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{k=1}^{n} \lambda_k^{2t} (\psi_k(i) - \psi_k(j))^2 \tag{8}$$

Aqui, como apresentado, λ são os autovalores de P e $\psi_k(i)$ as coordenadas de seus autovetores direitos. A prova para tal pode ser encontrada no material suplementar em [3].

O cálculo da distância de difusão deste modo é menos custoso computacionalmente e de fácil implementação. A partir do cálculo da matriz P, o algoritmo computa seus autovalores e autovetores e, dessa forma, consegue o mapeamento de acordo com as distâncias de difusão. O mapeamento pode, então, ser expresso em termos dos autovetores e autovalores de P como:

$$Z_i = \begin{bmatrix} \lambda_2^t \psi_2(i) \\ \lambda_3^t \psi_3(i) \\ \vdots \\ \lambda_n^t \psi_n(i) \end{bmatrix} \tag{9}$$

onde $\psi_2(i)$ indica o i-ésimo elemento do primeiro autovetor não constante de P associado ao maior autovalor diferente de 1, $\psi_3(i)$ o i-ésimo elemento do segundo autovetor de P associado ao segundo maior autovalor diferente de 1, e assim por diante. Os autovetores ortogonais direitos de P formam uma base para o espaço de difusão e os autovalores, por sua vez, associados a esses, fornecem a importância de cada dimensão. A redução de dimensionalidade pode ser atingida escolhendo m dimensões associados aos autovetores dominantes ($m < n$), o que permite que o conjunto de dados mapeados tenha menos parâmetros do que o conjunto original, fazendo $\left\|Z_j - Z_i\right\|$ se aproximar da distância de difusão, $D_t(\mathbf{x}_i, \mathbf{x}_j)$.

3 Resultados e discussões

O objetivo desta seção é a apresentação dos resultados obtidos com a execução do método exposto anteriormente na busca de *outliers* em dados reais experimentais expressos na forma de perfis unidimensionais provenientes de potenciostatos. Tais dados são os mesmos utilizados em [1] para o estudo da classificação usando a entropia de Tsallis. Dois aços inoxidáveis austeníticos comerciais foram expostos à uma rotina experimental na intenção de produzir dados de suas curvas de polarização. Tais curvas equivalem à parcela de corrente medida em diferentes faixas de potencial de 48 amostras de dois aços austeníticos comerciais, conhecidos como aços UNS S30400 e UNS S31600, doravante identificados como 304 e 316.

Neste experimento, as amostras foram expostas ao meio aquoso com 3,5% NaCl a 25 ºC e sob a taxa de varredura de potencial de 1,0 mV/s. Os perfis de corrosão das 24 amostras de cada aço são mostrados na Fig. 1.. Devido ao fluxo de elétrons, a corrente tem sinal negativo na região catódica e positiva na região anódica, invertendo-se dependendo da convenção de notação usada. Entretanto, utiliza-se o valor absoluto desta grandeza, e assim pode-se calcular o logaritmo da corrente. A principal diferença entre os aços é sua composição química, com o 316 contendo uma quantidade de cerca de 2,5% de molibdênio, contra um percentual residual encontrado no aço 304. Com esse

adicional, o aço 316 é significativamente mais resistente à corrosão nos meios contendo cloreto, principalmente em altos potenciais anódicos. Apesar de sua composição química diferente, as curvas se sobrepõem, principalmente nos potenciais relacionados à predominância de processos anódicos. É nessa faixa que fica visível a diferença entre os aços com relação à resistência à corrosão localizada junto com alguma dispersão natural. À primeira vista, é muito difícil atribuir uma determinada curva a um aço específico quando se tem todos agrupados. Nas curvas de polarização, utilizou-se o valor absoluto da corrente (A) e não a densidade de corrente (A/cm²).

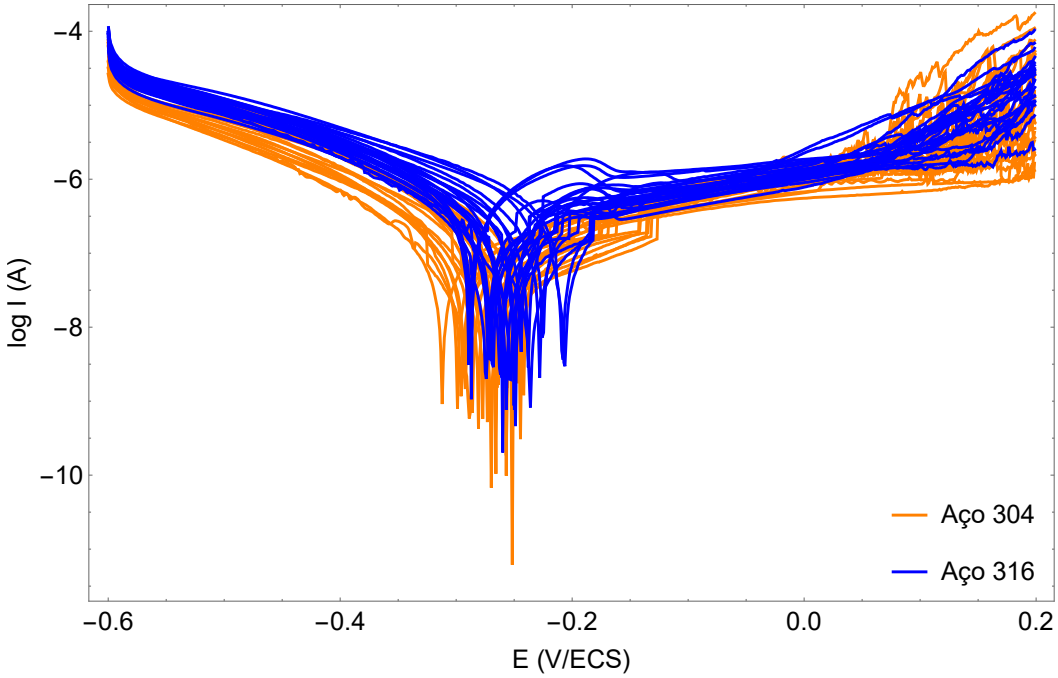


Figura 1: 48 curvas de polarização experimental dos aços inoxidáveis 304 (laranja) e 316 (azul). Cada curva consiste em 800 pontos de amostragem.

O primeiro passo na busca dos *outliers* da amostra de cada tipo de aço é a execução do algoritmo, assim como foi descrito na seção anterior, também para cada tipo de aço. O algoritmo foi executado considerando $\alpha = \varepsilon^2$, onde ε é o limite de conectividade do conjunto de dados ($\varepsilon = \min\{\|\mathbf{x}_i - \mathbf{x}_j\|, i \neq j, i, j = 1, \ldots, n\}$). Dado que objetiva-se encontrar os *outliers*, a matriz de difusão P deve ser o foco principal. O mapeamento dos perfis, a princípio, não é necessário.

Ao ordenar as colunas e linhas dessa matriz, segundo à ordem crescente (ou decrescente) das coordenadas do primeiro autovetor não constante, conseguimos uma maneira de agrupar os perfis, onde qualquer bloco (grande o suficiente) é considerado não-outlier. Na prática, ao exibir a matriz de difusão P na forma de um mapa de cores, perfis semelhantes são mostrados na forma de blocos de similaridade e, com isso, tais blocos representam *clusters* de acordo com o método. A justificativa da eficiência do uso deste autovetor na reordenação de um dado conjunto de perfis pode ser encontrada em [3].

A Fig. 2 mostra como o mapa de difusão unidimensional reorganiza revelando os perfis exibindo uma estrutura encoberta no conjunto. As matrizes de difusão dos aços 304 e 316 são vistas como um tipo de representação do conjunto de níveis de uma função $(i, j) \to (P)_{ij} \in \mathbb{R}$ com valor de P_{ij} representado por uma cor. Se P_{ij} está próximo de zero, a cor é branca e, quando P_{ij} está próximo de um, a cor é vermelha. Chamamos esta representação de mapa de cores da matriz de difusão. Na coluna à direita é possível observar um bloco de perfis semelhantes conectados pelo processo de difusão nos dois aços. Nas amostras do aço 304, primeiramente, é possível perceber um bloco que compreende os perfis de índice 6 a 23, no entanto, não tão uniforme quanto o bloco do aço 316 com os índices compreendidos de 4 a 18.

Para melhorar a visualização, podemos aumentar o tempo de difusão t, o que equivale a calcular as potências de P^t, possibilitando conectar melhor os perfis pelo processo de difusão (Fig. 3). Observe como agora os blocos de afinidade ficam bem mais claros com a mudança do parâmetro. Os blocos aparentes nos dois mapas de cores relativo às matrizes de difusão ordenadas indicam clusters conectados pelo processo de difusão. Conectados, os caminhos se formam ao longo de saltos curtos e de alta probabilidade. É possível observar também bloco menores, como no aço 316, com índices de 1 a 3, considerados, contudo, também *outliers* neste conjunto.

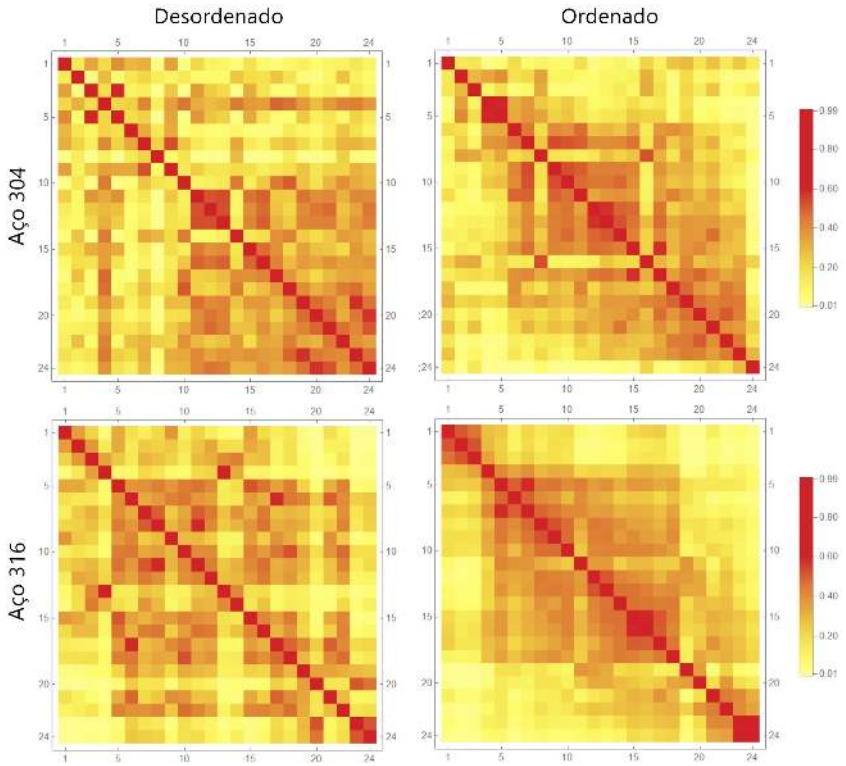


Figura 2: Representação das matrizes de difusão dos aços 304 (superior) e 316 (inferior) com $t = 1$.

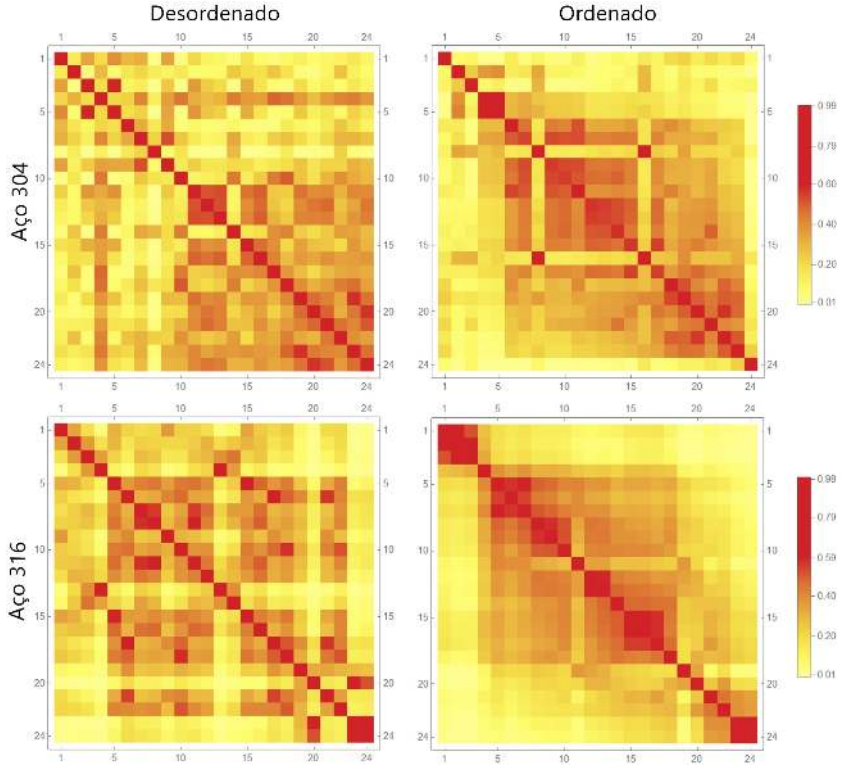


Figura 3: Mapa de cores dos aços 304 (superior) e 316 (inferior) agora com $t = 5$. O processo de difusão revela os *clusters*, permitindo assim que perfis semelhantes se conectem pela evolução do processo de difusão.

As inferências então podem ser realizadas. Referente ao aço 314, temos dois grupos: $R_1 = \{6, 7, 8, 9, 10, 11, 12, 13,$

$14, 15, 16, 17, 18, 19, 20, 21, 22, 23\}$, relativo ao maior bloco, e *outliers* $R_0 = \{1, 2, 3, 4, 5, 24\}$. Em relação ao aço 316, temos também dois grupos: $G_1 = \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}$, referente ao maior bloco, e *outliers* $G_0 = \{1, 2, 3, 19, 20, 21, 22, 23, 24\}$. Como observado, seguiu-se a hipótese de que qualquer bloco grande o suficiente é considerado não-outlier, caracterizando, assim, os demais perfis como discrepantes.

Ressaltamos que o índices apresentados pelo mapas de cores são os índices da amostra ordenada e não os índices da amostra original. Fazendo a associação adequada, na Tabela 1 temos o resultado proposto. Considerando que os experimentos tenham sido realizados de modo completamente randômico, pode-se calcular também a probabilidade de se ter amostras consecutivas de *outliers* no início ou fim do experimento, como duplas $(23, 24)$; triplas $(1, 2, 3)$ ou quádruplas $(1, 2, 3, 4)$ por análise combinatória. Como observado, os resultados propostos parecem demonstrar que houve uma dificuldade inicial e final com a padronização do experimento, evidenciado pelos números consecutivos de perfis indicados como *outliers* nos dois aços, principalmente nas amostras do aço 316.

Supondo que para a amostra do aço 304 tenham-se exatamente 6 *outliers*, a probabilidade das três primeiras observações estar entre eles é provavelmente baixa. De igual forma, supondo para o aço 316 exatamente 9 *outliers*, a probabilidade das 4 primeiras observações estarem entre os perfis discrepantes, dado que as duas últimas também estão, é ainda menos provável.

Calculando tais probabilidades, encontraram-se os valores aproximados de $0,988\%$ e $0,0624\%$ para as duas questões apresentadas, respectivamente. Pela baixa probabilidade em ambos os casos, pode-se atribuir os *outliers* a algumas fontes, tais como erro sistemático na fase de realização aprendizado dos ensaios (quando os métodos empregados não estavam sistematizados), a problemas de estacionaridade das medidas e/ou à dispersão dos parâmetros eletroquímicos.

Tabela 1: Classificação das curvas de polarização dos aços segundo análise do mapa de cores utilizando o algoritmo de mapas de difusão.

Aço	Não Outliers	Outliers
304	4, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 e 24	1, 2, 3, 5, 6 e 8
316	5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 18, 19, 21 e 22	1, 2, 3, 4, 13, 14, 20, 23 e 24

O mapeamento 1D ou 2D por meio dos mapas de difusão pode também ser utilizado para complementar a análise dos *outliers* por mapa de cores e ajudar a confirmar os perfis discrepantes pela disposição dos perfis mapeados. As Fig. 4 e Fig. 5 exibem o mapeamento 1D para ambos os aços. Em cada subfigura, a partir do mapeamento 1D da amostra dos aços contendo todos os 24 perfis, pode-se visualmente encontrar e excluir o perfil (ou perfis) mais afastado(s) dos demais e analisar o mapeamento para os que restaram. Apesar de ser uma análise subjetiva (que depende da análise visual), o procedimento pode ser válido como forma de complemento à análise pelo mapas de cores.

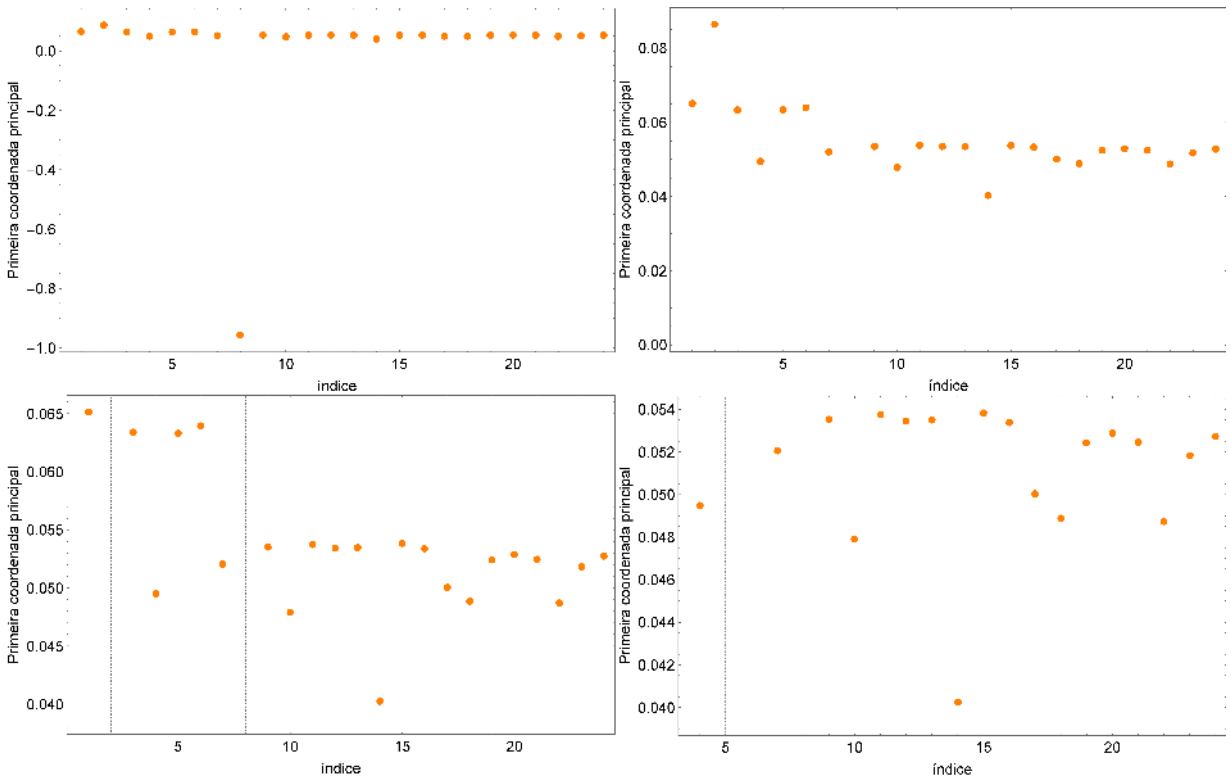


Figura 4: Mapeamento 1D para as amostras do aço 304. Em cada subfigura exclui-se o perfil mapeado mais distante.

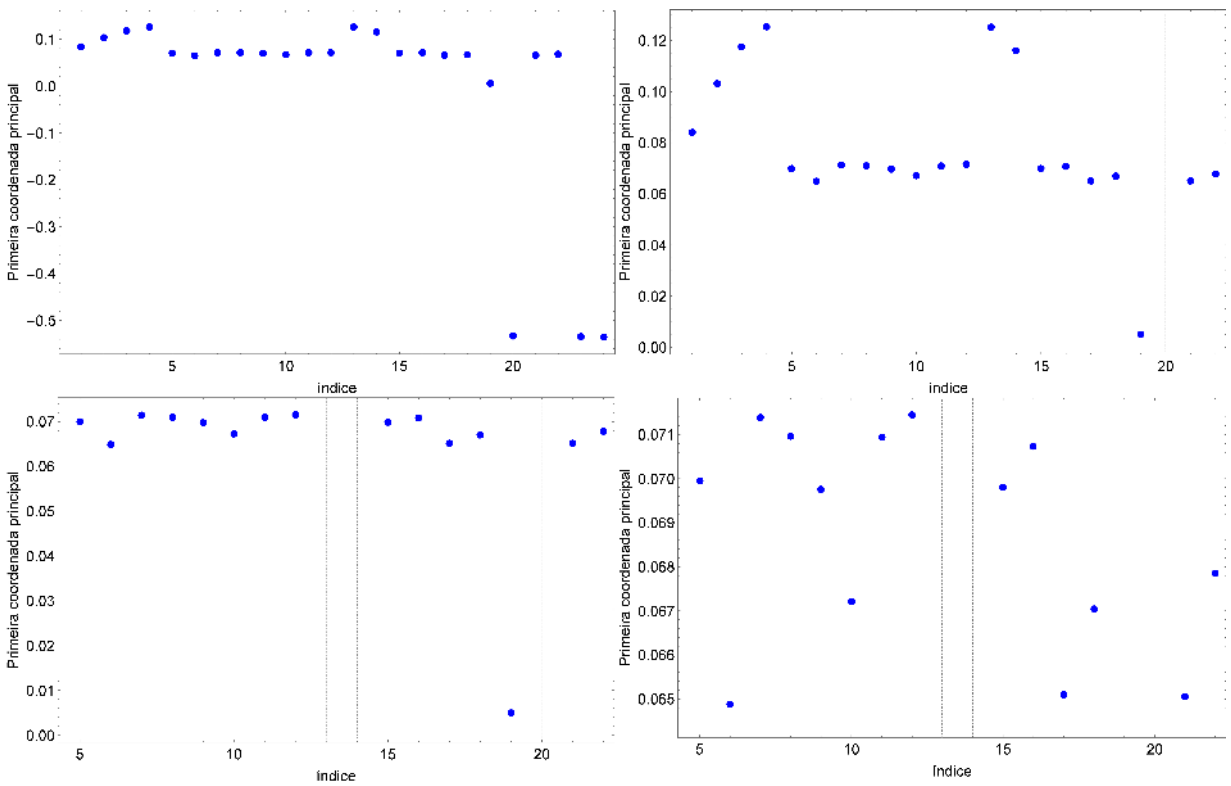


Figura 5: Mapeamento 1D para as amostras do aço 316. As subfiguras mostram o mapeamento a cada extração do possível *outlier*.

Na subfigura do canto superior esquerdo da Fig. 4, tem-se o mapeamento 1D com os 24 perfis do aço 304. A seguir, no canto superior direito, é excluído o perfil de nº 8 por estar mais distante. A subfigura no canto inferior esquerdo traz agora o mapeamento sem o perfil nº 2. Por fim, a última subfigura no canto inferior direito exclui os de números 1, 3, 5 e 6.

Em relação ao aço 316, a subfigura do canto superior esquerdo da Fig. 5 traz o mapeamento 1D com os 24 perfis agora do aço 316. No canto superior direito, são excluídos os perfis de nº 20, 23 e 24. A seguir, os de números 1, 2, 3, 4, 13 e 14 e, em seguida, o de nº 19.

Fazendo a comparação entre os *outliers* encontrados em ambas as abordagens (mapas de cores e visualização do mapeamento 1D), é possível ver que elas são coerentes, uma vez que apontam para o mesmo grupo. A Fig. 4 traz como *outliers* do aço 304 os perfis de nº 1, 2, 3, 5, 6 e 8. Por outro lado, a Fig. Fig. 5 traz como *outliers* do aço 316 também os perfis de nº 1, 2, 3, 4, 13, 14, 20, 23 e 24. Adicionalmente, o perfil nº 19 do aço 316 também foi considerado *outlier* na análise do mapeamento. É possível observar que ele também se encontra distante dos demais. No mapa de cores, entretanto, essa observação não era nítida.

As curvas de polarização, agora sem e com *outliers* de cada aço, são exibidas nas Fig. 4, 5 e 6.

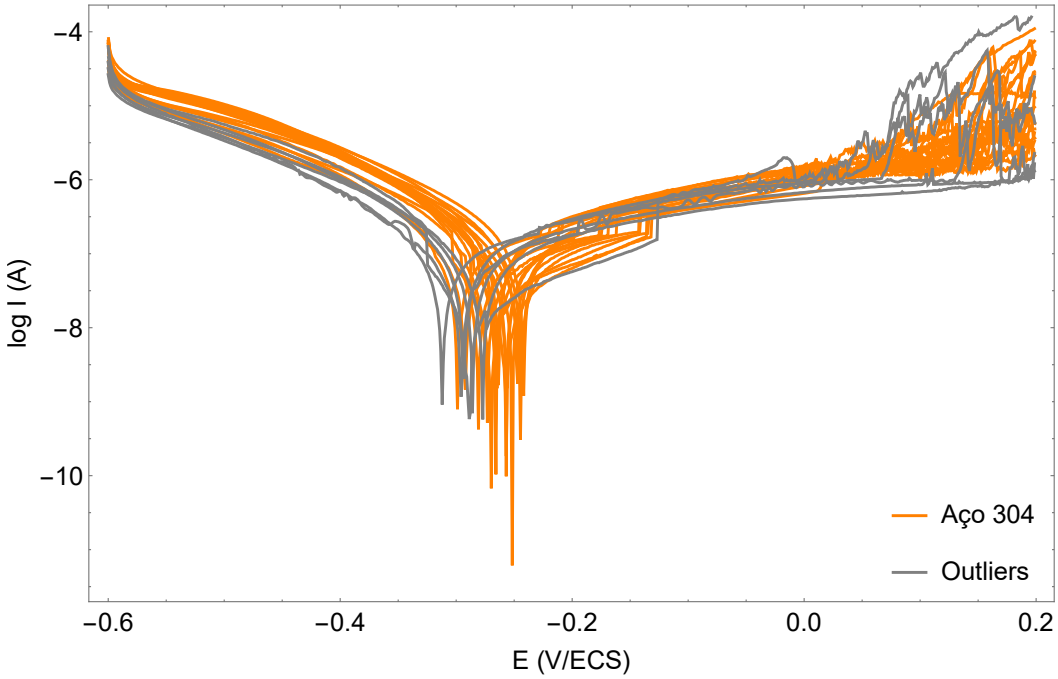


Figura 6: Curvas de polarização experimental do aço 304 com seus possíveis outliers.

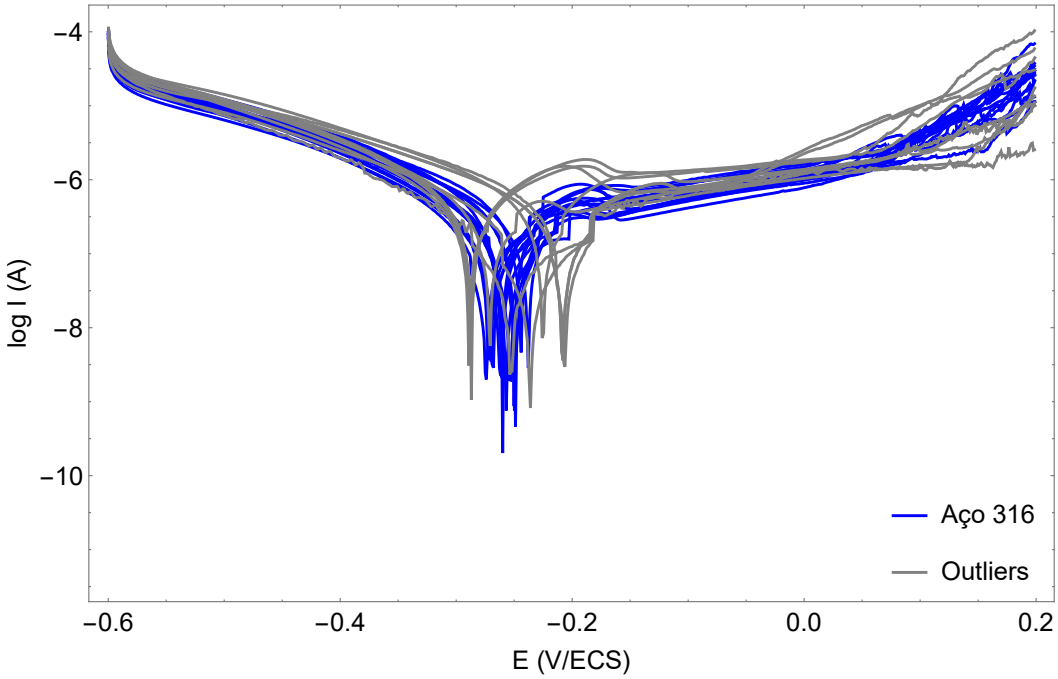


Figura 7: Curvas de polarização experimental do aço 316 com seus possíveis outliers.

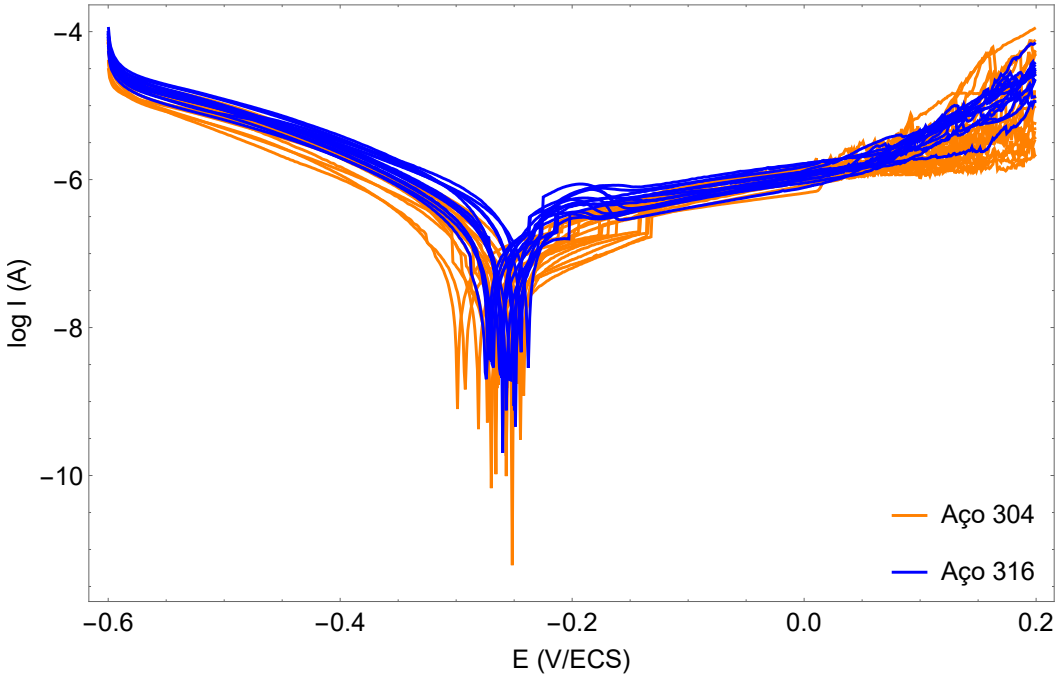


Figura 8: Curvas de polarização experimental do aços sem os possíveis outliers.

4 Conclusões

A técnica de mapas de difusão foi útil na identificação de *outliers* de um conjunto de dados experimentais expressos como perfis ao exibir os possíveis candidatos discrepantes por meio da matriz de difusão. Estes perfis correspondem a curvas de polarização potenciodinâmica amplamente empregadas em estudos de corrosão eletroquímica e, ao extrair os *outliers*, devem traduzir de forma adequada o comportamento do material imerso no meio corrosivo a

um dado potencial aplicado. Este resultado pode ser muito útil para o pesquisador da área na depuração dos dados experimentais.

Na amostra com 24 representantes do aço 304, seis foram classificados como *outliers*. Tratam-se dos perfis de nº 1, 2, 3, 5, 6 e 8. Em relação ao aço 316, foram dez: nº 1, 2, 3, 4, 13, 14, 19, 20, 23 e 24. Também foi possível concluir que, devido à baixa probabilidade de se obter perfis discrepantes sequencialmente no experimento, dado que foram observados outliers, tais representantes foram obtidos quando a rotina não estava ainda sistematizada e/ou o equipamento devidamente calibrado.

Em trabalhos futuros, a expectativa é a utilização dos resultados aqui apresentados na verificação do ganho na classificação em relação a todo o conjunto de dados. A identificação e eliminação dos *outliers* possibilita ao especialista que tome decisões tendo por base dados confiáveis, pré-selecionados por rigoroso procedimento numérico-matemático de mapas de difusão.

Agradecimentos

C.A.L.S. Oliveira agradece o suporte financeiro do Programa de Bolsa para Qualificação de Servidores – Instituto Federal do Norte de Minas Gerais. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Esta pesquisa contou ainda com o apoio financeiro da Faperj e do CNPq.

Referências

[1] R. Fabbri, I. N. Bastos, F. D. M. Neto, F. J. Lopes, W. N. Gonçalves, e O. M. Bruno, "Multi-*q* pattern classification of polarization curves," *Physica A: Statistical Mechanics and its Applications*, vol. 395, pp. 332–339, 2014. Disponível em: https://doi.org/10.1016/j.physa.2013.09.048

[2] G. Wang, J. Liu, Y. Li, e L. Shang, "Fault detection based on diffusion maps and *k* nearest neighbor diffusion distance of feature space," *Journal of Chemical Engineering of Japan*, vol. 48, no. 9, pp. 756–765, 2015. Disponível em: https://doi.org/10.1252/jcej.14we227

[3] F. M. Neto, P. Souza, e M. de Magalhães, "Determining baseline profile by diffusion maps," *European Journal of Operational Research*, vol. 279, no. 1, pp. 107–123, 2019. Disponível em: https://doi.org/10.1016/j.ejor.2019.05.032

[4] R. R. Coifman e S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006. Disponível em: https://doi.org/10.1016/j.acha.2006.04.006

[5] R. R. Coifman e M. J. Hirn, "Diffusion maps for changing data," *Applied and Computational Harmonic Analysis*, vol. 36, no. 1, pp. 79–107, 2014. Disponível em: https://doi.org/10.1016/j.acha.2013.03.001

[6] S. S. Lafon, "Diffusion maps and geometric harmonics," Tese de doutorado, Yale University, New Haven, Estados Unidos, 2004. Disponível em: https://sites.google.com/site/stefansresearchpapers/