Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto Politécnico

Carlos Alberto Lopes dos Santos de Oliveira

Análise de sinais eletroquímicos por mapas de difusão

Nova Friburgo
2022

Carlos Alberto Lopes dos Santos de Oliveira

Análise de sinais eletroquímicos por mapas de difusão

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Modelagem Computacional, da Universidade do Estado do Rio de Janeiro.

Orientadores: Prof. Francisco Duarte Moura Neto, Ph.D.
Prof. Ivan Napoleão Bastos, D.Sc.

Nova Friburgo
2022

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/E

O48 Oliveira, Carlos Alberto Lopes dos Santos de Oliveira.
Análise de sinais eletroquímicos por mapas de difusão / Carlos Alberto Lopes dos Santos de Oliveira. - 2022.
167 f. : il.

Orientador: Francisco Duarte Moura Neto.
Orientador: Ivan Napoleão Bastos.
Tese (doutorado) – Universidade do Estado do Rio de Janeiro, Instituto Politécnico.

1. Impedância– Teses. 2. Difusão – Modelos matemáticos – Teses. I. Moura Neto, Francisco Duarte. II. Bastos, Ivan Napoleão III. Universidade do Estado do Rio de Janeiro. Instituto Politécnico. IV. Título.

CDU 537.311.6

Bibliotecária Pâmela Lisboa CRB7/ 5965

_____ _____

Assinatura Data

Carlos Alberto Lopes dos Santos de Oliveira

Análise de sinais eletroquímicos por mapas de difusão

Tese apresentada como requisito parcial para obtenção do título de Doutor ao Programa de Pós-Graduação em Modelagem Computacional do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

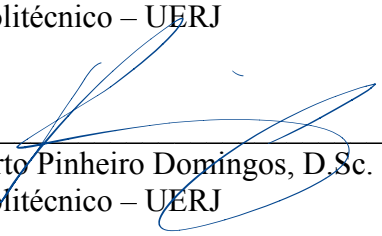
Aprovada em 22 de julho de 2022

Banca examinadora:

Prof. Francisco Duarte Moura Neto, D.Sc. (Orientador)
Instituto Politécnico – UERJ

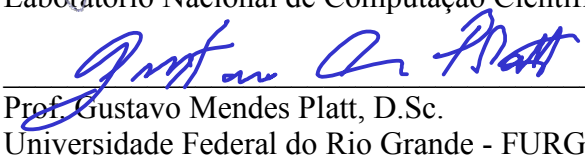
Prof. Ivan Napoleão Bastos, D.Sc. (Orientador)
Instituto Politécnico – UERJ

Prof. Bernardo Sotto Maior Peralva D.Sc.
Instituto Politécnico – UERJ



Prof. Roberto Pinheiro Domingos, D.Sc.
Instituto Politécnico – UERJ

Prof. Gustavo Barbosa Libotte, D.Sc.
Laboratório Nacional de Computação Científica - LNCC



Prof. Gustavo Mendes Platt, D.Sc.
Universidade Federal do Rio Grande - FURG

Nova Friburgo

2022

DEDICATÓRIA

Este trabalho é dedicado a todos que, assim como eu, são sonhadores, àqueles que sonham com um mundo mais justo e com mais oportunidades, àqueles que acreditam que sem a fé não somos nada, àqueles que acreditam que conquistas pessoais são importantes, porém mais importante ainda é o que fazemos em prol do outro, àqueles que procuram ser melhores a cada dia, para si e para outrem.

AGRADECIMENTOS

Agradeço, em especial, aos professores Francisco Moura Neto e Ivan Bastos, responsáveis pela introdução ao tema, orientação, motivação e revisão na realização deste trabalho.

Agradeço também à instituição UERJ e a todos os professores que me acompanharam durante o período dedicado ao curso de pós-graduação. A saber, professora Daiara Fernandes e professores Luiz Abreu, Germano Monerat, Diego Knupp, Bernardo Peralva e Ricardo Fabbri. Cada um contribuiu consideravelmente para minha formação compartilhando suas experiências e sabedoria. Agradeço também ao professor Odemir Bruno da USP pelo compartilhamento de informações que possibilitaram construir parte da seção de resultados desse trabalho.

Agradeço a Deus, mormente, pela força e coragem durante toda esta longa caminhada.

À minha companheira Jéssica, pessoa com quem amo partilhar a vida. Obrigado pelo carinho, pela paciência e por sua capacidade de me trazer paz na correria desse mundo louco.

Aos meus pais pelo amor incondicional, meus irmãos pelo carinho e respeito e aos meus filhos Mateus, Manuela e Marcela que constituem as maiores riqueza em minha vida.

Aos amigos que fiz durante o período de curso e, em especial, ao amigo Taciano por todo incentivo, apoio e disponibilidade.

E finalmente o reconhecimento ao IFNMG pelo suporte concedido por meio do programa PBQS para o desenvolvimento deste trabalho.

Aquilo que não sei, também não penso saber.

Sócrates

RESUMO

OLIVEIRA, Carlos Alberto Lopes dos Santos de. *Análise de sinais eletroquímicos por mapas de difusão*. 2022. 167 f. Tese (Doutorado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2022.

Esta tese propõe um uso inovador no estudo de sinais eletroquímicos pela técnica de redução de dimensionalidade não linear conhecida como mapas de difusão. Os mapas de difusão conseguem descobrir padrões em conjuntos de dados, agrupando-os e considerando diversas escalas para classificação. A redução de dimensionalidade foi aplicada inicialmente a imagens digitais, mostrando a capacidade de organizar objetos e revelando a potencialidade desta técnica para tratar dados. Além de fornecer uma revisão bibliográfica geral do método e explicar detalhadamente seu funcionamento, no trabalho analisa-se de modo abrangente o comportamento qualitativo da técnica em termos de parametrização da modelagem, trazendo para o debate questões ainda pouco compreendidas sobre o método. A seguir, é apresentada uma aplicação para classificação de perfis de sinais eletroquímicos de curvas de polarização e de espectroscopia de impedância eletroquímica, utilizados principalmente para avaliar a cinética de corrosão de eletrodos em ambientes corrosivos. A análise utilizou os perfis de testes experimentais com dezenas de réplicas. A classificação correta é um desafio, pois a forte não linearidade dos perfis, além da sobreposição, torna-se uma tarefa difícil. Além disso, deve-se imaginar que *outliers* estão presentes e estes podem ter origens diversas, como variações de materiais, erros experimentais e artefatos, e o procedimento de classificação deve ser sensível para detectá-los a partir do conjunto de dados. Os resultados obtidos são considerados promissores, atingindo-se taxas de 81, 88 e 79% para as faixas de potencial eletroquímico baixo, alto e em toda a faixa estudada. Estes resultados são comparados com métodos clássicos de classificação em termos de eficiência atingida. Este trabalho também mostra como os mapas de difusão podem ser úteis na busca de *outliers* de sinais eletroquímicos. Após depurar os dados experimentais com a técnica, o classificador atinge 94% na faixa de alto potencial, quando o melhor resultado é obtido. Por fim, a tese também apresenta uma análise crítica do efeito do parâmetro de escala que mede a similaridade entre os dados do mapeamento obtido com os mapas de difusão utilizando dados simulados. Esta parametrização se refere à conectividade estrutural e como os dados podem ser conectados.

Palavras-chave: Redução de dimensionalidade. Mapas de difusão. Outliers. Curvas de polarização. Impedância eletroquímica. Imagens digitais.

ABSTRACT

OLIVEIRA, Carlos Alberto Lopes dos Santos de. *Analysis of electrochemical signals by Diffusion Maps*. 2022. 167 f. Tese (Doutorado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2022.

This thesis proposes an innovative use in the study of electrochemical signals by the technique of nonlinear dimensionality reduction known as Diffusion Maps. Diffusion maps can discover patterns in datasets, grouping them and considering different scales for classification. The dimensionality reduction was initially applied to digital images, showing the capacity of organizing objects and revealing the potentiality of this technique to treat data. In addition to providing a general bibliographic review of the method and explaining its operation in detail, the work comprehensively analyzes the qualitative behavior of the technique in terms of modeling parameterization, bringing to the debate issues that are still poorly understood about the method. Next, an application for classifying electrochemical signal profiles from polarization curves and electrochemical impedance spectroscopy is presented, mainly used to evaluate the corrosion kinetics of electrodes in corrosive environments. The analysis used the profiles of experimental tests with dozens of replicates. Correct classification is a challenge, as the strong non-linearity of the profiles, in addition to overlap, becomes a difficult task. Furthermore, one must imagine that outliers are present and these may have different origins, such as material variations, experimental errors and artifacts, and the classification procedure must be sensitive to detect them from the dataset. The results obtained are considered promising, reaching rates of 81, 88 and 79% for the low and high electrochemical potential ranges and for the entire studied range. These results are compared with classical classification methods in terms of efficiency achieved. This work also shows how Diffusion Maps can be useful in the search for outliers of electrochemical signals.After debugging the experimental data with the technique, the classifier reaches 94% in the high potential range, when the best result is obtained. Finally, the thesis also presents a critical analysis of the effect of the scale parameter that measures the similarity between the mapping data obtained with the diffusion maps using simulated data. This parameterization refers to structural connectivity and how data can be connected.

Keywords: Dimensionality Reduction. Diffusion Maps. Outliers. Polarization curves. Electrochemical impedance. Digital images.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

#PCC	número de perfis classificados corretamente
isomap	*isometric mapping* - mapas de características isométricas
LE	*laplacian eingenmaps* - automapas do laplaciano
LLE	*locally linear embedding* - imersão localmente linear
KPCA	*kernel principal component analysis* - análise de componentes principais do núcleo
MDS	*multidimensional scaling* - escalonamento multidimensional
NLDR	*non-linear dimension reduction* - redução de dimensão não-linear
PCA	*principal component analysis* - análise de componentes principais
RGB	sistema de cores *Red-Green-Blue* - vermelho-verde-azul
RD	redução de dimensionalidade
RSS	*residual sum of squares* - soma residual dos quadrados
SGE	*semi-group error* - erro do semigrupo
Tc	taxa de acertos do classificador

LISTA DE SÍMBOLOS LATINOS

a	Limite inferior do intervalo de α
A	Matriz de dados ajustados para a PCA
A_1	Grupo considerado representativo para os perfis de impedância em relação à variável módulo
A_2 e A_3	Grupos considerados *outliers* para os perfis de impedância em relação à variável módulo
b	Limite superior do intervalo de α
B	Matriz cujas colunas contém os m autovetores associados aos autovalores dominantes da matriz R
B_1	Grupo considerado representativo para os perfis de impedância em relação à variável fase
B_2	Grupo considerado *outlier* para os perfis de impedância em relação à variável fase
B_{304}	Grupo tomado representativo para os perfis de corrosão relativo ao aço 304
B_{316}	Grupo tomado representativo para os perfis de corrosão referente ao aço 316
c	Fator de ponderação da função do mapeamento não-linear de *Sammon*
C_1	Grupo considerado representativo para os perfis de impedância em relação à variável módulo e fase
C_2	Grupo considerado *outlier* para os perfis de impedância em relação à variável módulo e fase
d	Número de dimensões do espaço de alta dimensão (*feature space*)
D_i	i-ésimo dígito siginificativo
D_X	Matriz de distâncias iniciais do MDS
D_Z	Matriz de distâncias no espaço reduzido
E	Matriz peso do LLE
$\mathscr{E}$	Conjunto de arestas do grafo
$f(Z)$	Função custo do LE
$\mathbf{f}$	Autovetores da matriz laplaciana
G	Matriz diagonal com a soma das entradas da matriz W
$\mathscr{G}$	Grafo
H^*	Matriz de *Gram* do LLE
I_n	Matriz identidade $n \times n$
J	Matriz centralizadora do MDS
$\tilde{k}$	Núcleo de difusão
$\tilde{k}_{ij}$	Entradas da matriz $\tilde{K}$
$\tilde{K}$	Matriz de similaridades dos mapas de difusão
k_{ij}	Entradas da matriz K

K	Matriz de similaridades ajustada dos mapas de difusão
K^*	Matriz K normalizada
L	Matriz laplaciana do grafo
$\mathscr{L}$	Matriz laplaciana normalizada
m	Número de dimensões do espaço reduzido
$M(\alpha)$	Função de variação global
n	Número de pontos do espaço original
O_{304}	Grupo tomado *outlier* para os perfis de corrosão relativo ao aço 304
O_{316}	Grupo tomado *outlier* para os perfis de corrosão referente ao aço 316
p_{ij}	Entradas da matriz P
P	Matriz de difusão
P^t	Potência da matriz P de expoente t
P^t_{ij}	Soma dos elementos da linha i da matriz P^t
q	Número de vizinhos próximos no LLE
R	Matriz de covariância dos dados de entrada para a PCA
Q	Matriz de autovetores direitos da matriz P como colunas
s	Valor arbitrário para o i-ésimo dígito siginificativo
S	Matriz de autovetores da matriz K^* como colunas
t	Parâmetro temporal nos mapas de difusão
T	Parâmetro de escala do *kernel* gaussiano utilizado no LE
tr	Traço de uma matriz
$\mathbf{u}$	Autovetores da matriz laplaciana normalizada
U	Matriz de autovetores da matriz laplaciana normalizada como colunas
v_i	Vértices do grafo
V_n	Conjunto de vértices do grafo
$\mathbf{v}$	Vetor direção unitário para a PCA
W	Matriz de similaridades do LE
$\mathbf{x}_i$	Pontos (vetores) de dados do espaço de alta dimensão
X	Conjunto de pontos de dados do espaço de alta dimensão
Y	Matriz diagonal com a soma das entradas da matriz K
$\mathbf{z}_i$	Pontos (vetores) de dados do espaço reduzido
$\|z_i\|$	Módulo de um sinal i de impedância
Z	Conjunto de pontos de dados do espaço reduzido
$\mathbf{1}$	Vetor de uns com comprimento n

LISTA DE SÍMBOLOS GREGOS

α	Parâmetro de escala do núcleo gaussiano
γ	Ruído aleatório gaussiano adicionado aos perfis simulados
$\delta_i(l)$	l-ésima componente do i-ésimo autovetor da matriz de covariância R
Δ	Máxima diferença percentual entre as probabilidades dos primeiros dígitos significativos de uma distribuição dada e a distribuição de Benford
ϵ	Parâmetro limite de conectividade do grafo
ζ_j	Produto interno entre o vetor direção unitário para a PCA e uma coluna j da matriz A de dados ajustados
ε_{m}	Limite de conectividade do conjunto de dados
ε_{d}	Diâmetro do conjunto de dados
$\varepsilon_{\mathrm{mM}}$	Parâmetro de escala *min-max* do conjunto de dados
$\varepsilon_{\mathrm{mean}}$	Parâmetro de escala médio do conjunto de dados
θ	Autovalor da matriz laplaciana
λ	Autovalor da matriz de difusão
λ^t	Potência do autovalor λ de expoente t
Λ	Matriz diagonal de autovalores da matriz de similaridade normalizada K^*
μ	Multiplicador de Lagrange para a função custo de reconstrução do LLE
Ξ	Matriz cujos autovalores e autovetores dão origem às coordenadas dos dados mapeados no LLE
ϖ	Parâmetro de regularização de *Thikonov* do LLE
ς	Multiplicador de Lagrange para a função custo da PCA
ϱ	Multiplicador de Lagrange para a função custo do LLE
ϕ_l	Autovetores da matriz de similaridade normalizada K^*
$\mathbf{\Phi}$	Função reconstrução de dados no LLE
$\psi(i)$	Coordenada i do autovetor direiro da matriz de difusão
$\boldsymbol{\psi}$	Autovetores direitos da matriz de difusão
$\mathbf{\Psi}$	Função mapa de difusão
ω_l	Autovetores esquerdos da matriz de difusão P

SUMÁRIO

INTRODUÇÃO

Motivação

A chamada aprendizagem de máquinas consiste, muitas vezes, em desenvolver um classificador ou estimar uma função, a partir de um conjunto de dados. Em várias dessas situações os dados apresentam muitas características numéricas. No entanto, há razões para considerar que tais informações pertençam a uma variedade de baixa dimensão, ou seja, esses atributos podem ser dependentes entre si, como no caso de variáveis correlacionadas ou, de outra forma, a dimensionalidade dos dados pode ser maior do que a necessária. Isso, em outras palavras, significa dizer que talvez a verdadeira geometria intrínseca do conjunto de dados, aquela onde somente as variáveis importantes residem, possa ser obtida com um número menor de parâmetros.

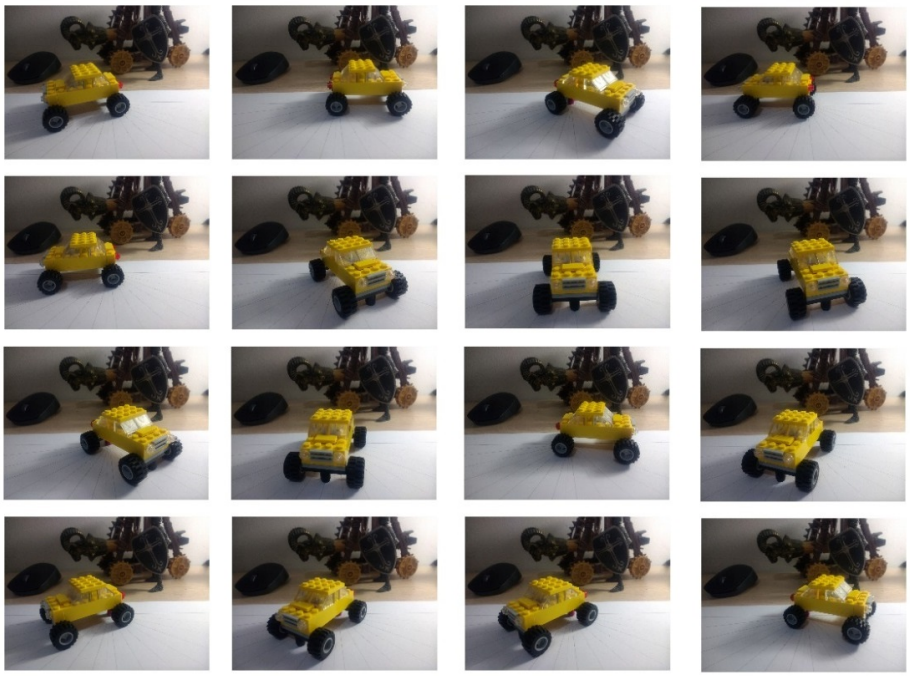
De acordo com Dias (2012), o estudo de dados de alta dimensionalidade é inevitável e possui motivação prática e teórica. Do ponto de vista prático, procura-se analisar esses dados multivariados buscando modelar os fenômenos que os geraram na esperança de entender sobre seu funcionamento. Do ponto de vista teórico, deve-se considerar a chamada "maldição da dimensionalidade" (*curse of dimensionality*), um termo que remete aos problemas associados ao processamento de dados de alta dimensão (problemas presentes na atual sociedade orientada pela informação). Neste contexto, os dados de dimensionalidade elevada são esparsos no espaço de alta dimensão. Em Lee e Verleysen (2007) podem ser verificadas algumas propriedades inusitadas relacionadas a um espaço de alta dimensão, como por exemplo em relação a normas e distâncias entre os dados.

Imagine, por exemplo, uma imagem em tons de cinza de 100×100 píxeis de um rosto humano, no qual cada píxel pode ser visto como uma representação de uma variável, uma tonalidade de cinza em uma posição específica da imagem. Logo, tem-se um espaço de 10.000 dimensões. Dado um espaço de características com tão alta dimensão, os pontos de dados amostrados, um conjunto de imagens, normalmente aparecem dispersos nesse espaço. A desvantagem, então, é que, ao tentar modelar tais fenômenos que geraram tais dados ou buscar alguma informação específica presente em uma ou mais imagens, alguns algoritmos distorcem ou falham por completo na estimativa da função. Há ainda que considerar que os dados possam estar corrompidos por ruídos, ou o modelo de geração dos dados pode não comportar determinados atributos como, por exemplo, um bigode, uma máscara, um tapa olho, um óculos, etc, ou até mesmo o reconhecimento de que na imagem há uma pessoa. Isso acontece, pois, de acordo com Lafon (2004), a estimativa da função e da densidade torna-se dispendiosa ou imprecisa, e as medidas de similaridade global desfazem-se ou perdem significado.

A busca pela noção adequada de similaridade comum entre pontos de dados mede

sua organização eficiente, o que, por sua vez, tem vantajoso impacto nas inferências também no campo do reconhecimento de padrões. Considere, por exemplo, uma coleção de imagens de um brinquedo infantil (Fig. 1). Na figura temos 16 imagens de um mesmo brinquedo com cada imagem com 400 por 300 píxeis. Em princípio, o ângulo de rotação do brinquedo é o único grau de liberdade no conjunto de imagens, mas há outros, por exemplo, com a translação do carrinho.

Figura 1 - Brinquedo infantil rotacionado com ângulo aleatório



Fonte: O autor, 2022.

Uma pessoa, diante da tarefa de organizar as imagens, muito provavelmente notaria que elas diferem pela rotação do brinquedo e atribuiria a este parâmetro a maior importância. Com isso, ordenaria as imagens sem dificuldade. Um computador, por outro lado, identificaria cada imagem como um ponto de dados em $\mathbb{R}^{400\times300\times3}$, um espaço de coordenadas de alta dimensão, e o expoente 3 se explica por ser uma imagem colorida, em que se utiliza os sistema de cores RGB. Na busca pela organização automática das imagens, os pontos de dados seriam, por natureza, organizados de acordo com sua posição no espaço coordenado, cuja medida de similaridade mais comum é relacionada à distância euclidiana, de forma inversa. Quanto menor essa distância, maior a similaridade, e vice-versa. Organizando-se as imagens de acordo com as distâncias entre cada coordenada dos pontos de dados, decerto, o computador, por meio de uma técnica eficaz, também conseguiria realizar a tarefa.

A ideia subjacente à realização da tarefa está na hipótese de que a dimensionalidade

intrínseca das imagens—as características que as individualizam—então, pode ser muito menor. De fato, nesse exemplo, o principal parâmetro pertinente para organizá-las é o ângulo de rotação do brinquedo. Em Belkin (2003), é citado um exemplo relacionado à fala—enquanto representações típicas de sinais de fala são baseadas em transformadas de Fourier de janela deslizante (uma particularidade da transformada de Fourier) e são de alta dimensionalidade, o próprio sinal de fala é produzido pelo trato vocal, que tem um grau de liberdade limitado.

No mapeamento do espaço onde os dados residem, uma pequena distância euclidiana entre vetores quase certamente indica que eles são muito similares. Contudo, uma grande distância, em contrapartida, fornece pouca informação sobre a natureza da discrepância. Eis a dificuldade: nos espaços onde naturalmente os dados residem, que são de alta dimensão, as distâncias podem variar enormemente. A distância euclidiana, portanto, fornece apenas uma boa medida de similaridade local entre os pontos com esta pequena distância relativa.

A saída, então, para a redução da dimensionalidade não-linear, é a hipótese de que os dados estão sobre ou ao redor de uma variedade de baixa dimensão em um espaço dimensional (provavelmente) muito elevado. A obtenção de informações qualitativas sobre essa variedade diferenciável hipotética é denominada de aprendizagem de máquina (*manifold learning*, do inglês).

Dessa forma, quando a hipótese de o conjunto de dados ser uma amostra de uma variedade diferenciável fizer sentido, o ideal seria se fosse possível medir distâncias na própria variedade em vez de no espaço euclidiano e com isso mapear os dados e a relação entre pontos individuais usando menos dimensões. Nas imagens do brinquedo (Fig. 1), por exemplo, levando em conta sua estrutura global, pode-se representar os dados do conjunto de imagens usando apenas uma variável: o ângulo de rotação do brinquedo.

Outro exemplo de conjunto de dados de alta dimensão que pode conter informações representativas escondidas em cada uma das suas entradas são os provenientes de experimentos envolvendo sinais eletroquímicos. Uma curva de polarização de um aço inoxidável, por exemplo, é um gráfico de densidade de corrente *versus* potencial que pode ser associado a um vetor com múltiplas entradas contendo tantas dimensões quantas forem os intervalos de medida de potencial dentro de uma faixa específica. Por outro lado, como estas diferentes curvas estão associadas às quantificações do comportamento destes aços inoxidáveis expostos a várias condições onde a diferença se dá principalmente na faixa de alto potencial, é razoável imaginar que cada uma destas curvas caracterize o comportamento de certo aço exposto a dada rotina experimental, principalmente em faixas específicas. Isto posto, talvez estes sinais possam ser representados com menos informações numéricas explicativas.

O objetivo, então, da redução de dimensionalidade, é determinar uma estrutura de dados de menor dimensão que os represente, levando a um mapeamento significativo. Tal

representação reduz a dimensionalidade, enquanto, idealmente, deve preservar as relações importantes entre os pontos de dados.

Esta tese explora um método para a realização desse propósito, que são os mapas de difusão (*Diffusion Maps*), aplicando-os a dados físicos da área de corrosão e eletroquímica. Além desta contribuição, o trabalho explora também questões ainda não completamente resolvidas na literatura, nomeadamente, a escolha dos parâmetros mais significativos dos mapas de difusão, com o objetivo de melhorar a representação dos dados em um espaço de dimensão reduzida.

Objetivos principais da tese

O principal objetivo desta tese é avaliar a técnica de mapas de difusão na eficiente redução de dimensionalidade e depuração em perfis provenientes de sinais eletroquímicos de curvas de polarização e de espectroscopia de impedância eletroquímica. Como contribuição original adicional, é apresentado um procedimento útil para o entendimento aprofundado do parâmetro de escala no núcleo de similaridade, na técnica de mapas de difusão, e sua escolha.

Objetivos específicos

Fazendo uso da técnica de mapas de difusão, são delineados os seguintes objetivos específicos:

- Apresentar uma revisão bibliográfica geral sobre a técnica e explicar o seu funcionamento;
- Discutir os parâmetros envolvidos em seu algoritmo e sua sensibilidade relativamente a estes diante do caso de estudo, no mapeamento, buscando a redução de dimensionalidade;
- Comparar o método com outras estratégias presentes na literatura na busca de uma redução de dimensionalidade significativa e eficiente;
- Avaliar *outliers* por meio da técnica de mapas de difusão, ampliando seu uso também em relação à depuração dos dados;
- Reavaliar o ganho da classificação eficiente com a redução de dimensionalidade após a depuração dos dados;

- Estudar o efeito do parâmetro de escala para os resultados obtidos com a técnica em perfis gerados artificialmente.

Organização do trabalho

A presente tese está dividida em cinco capítulos. Inicialmente, o capítulo 1 traz a revisão bibliográfica da técnica de mapas de difusão, um exemplo motivacional envolvendo imagens digitais e uma breve apresentação de alguns dos principais métodos para redução de dimensionalidade. O objetivo de relacioná-los é apresentar um panorama sobre os principais métodos fisicamente motivados existentes e discorrer sobre suas características.

No capítulo 2 é apresentada a técnica de mapas de difusão. Por meio de uma linguagem simples e objetiva, a expectativa é que o leitor compreenda como e por que a técnica funciona e, com isso, consiga entender sua aplicação e suas limitações nos casos de estudo apresentados.

Já o capítulo 3 apresenta uma aplicação dos mapas de difusão para um caso específico no aprendizado de máquinas de sinais eletroquímicos de curvas de polarização de perfis de dois tipos de aços inoxidáveis austeníticos em meio aquoso contendo cloreto. O objetivo aqui é comparar o mapeamento obtido com os mapas de difusão e os resultados da classificação supervisionada com outros diferentes métodos da literatura. Adicionalmente, são apresentadas análises em faixas distintas de potencial utilizando os métodos propostos e os resultados são também comparados com outras técnicas clássicas discutidas no capítulo 1.

O capítulo 4 aborda a aplicação dos mapas de difusão na busca de *outliers* de sinais eletroquímicos. Inicialmente, mostra-se como conseguir a depuração dos dados e, em seguida, a metodologia utilizada no capítulo 3 é novamente aplicada agora a esse novo conjunto de perfis. Os resultados com e sem *outliers* são analisados. Além disso, é feita também a busca de *outliers* em sinais de impedância eletroquímica de uma nova amostra para outro aço comercial, o aço UNS S32750.

O capítulo 5 traz um estudo aprofundado sobre o efeito do parâmetro de escala da técnica em dados artificiais. Tal análise é extremamente relevante, uma vez que a escolha deste parâmetro na literatura, em geral, não é muito clara. Mostra-se que variadas escolhas desse parâmetro implicam em diferentes resultados com a técnica em relação ao mapeamento obtido e, desta forma, é sugerido uma nova maneira de escolher o parâmetro baseada na construção de uma medida global do mapeamento e na sua variação, que definem distintas zonas de funcionamento qualitativo dos mapas de difusão.

Por fim, são apresentadas as conclusões deste trabalho e as perspectivas para a continuação desta pesquisa. O anexo A traz o algoritmo básico da PCA, o anexo B o algoritmo básico do LLE, o anexo C trata da função custo do LE e como ela pode ser

escrita na forma matricial, o anexo D traz a justificativa de equivalência entre a distância de difusão no espaço de recurso e a distância euclidiana no espaço de difusão e o anexo E apresenta os trabalhos anteriormente publicados com parte dos resultados desta pesquisa.

1 REVISÃO BIBLIOGRÁFICA

Neste capítulo é apresentada uma breve revisão sobre os principais assuntos que motivaram este trabalho, a saber, redução de dimensionalidade, métodos usuais para redução, mapas de difusão e sinais eletroquímicos. Apesar de procurar apresentá-los na ordem descrita, para fins de organização, isto não quer dizer que não estejam correlacionados. De fato, estes têm conexões dentro do universo da análise de dados de alta dimensão e o objetivo deste capítulo é mostrar as principais contribuições já realizadas por diversos autores acerca dos tópicos descritos. Como ponto de partida, o tema comum que envolve esses temas é a análise de dados.

Para começar, assume-se que um dado, de modo simples, compreende uma coleção numérica de valores que registram a magnitude de vários atributos no objeto de estudo (BERTHOLD, 2007). A análise de dados, por sua vez, compreende o processamento desses dados, para determinar eventuais características implícitas nestas informações.

Ao procurar entender o universo à nossa volta, o ser humano coleta e registra informações há milhares de anos. A necessidade de compreensão do mundo que nos cerca sempre foi motivada por questões muito práticas que se confundem, quase sempre, com o próprio instinto de sobrevivência. A invenção dos sistemas de escrita, por exemplo, tem pelo menos 5 mil anos (ROBINSON, 2018) e os registros astronômicos mais antigos datam dessa época e se devem aos chineses, babilônios, assírios e egípcios. Esses registros permitiram a preservação da língua sobre os objetos materiais e, sem dúvidas, foram cruciais na evolução cultural e sobrevivência dos diferentes povos que habitam o nosso planeta.

Apesar da humanidade procurar organizar um saber rigoroso desde a Grécia antiga (século VII a.C.), aspirando um conhecimento racional sobre o mundo em oposição ao mito e ao saber comum, o nascimento do conhecimento científico só é considerado com Galileu Galilei ($1564 - 1642$) e tem pouco menos de 400 anos. Esse conhecimento racional, durante toda a Antiguidade e Idade Média, foi chamado de filosofia e compreendia diversos tipos de conhecimento que se estendiam por áreas distintas.

Com o advento da ciência moderna, nasce a determinação de um objeto específico de investigação e, com isso, a necessidade de métodos pelos quais se farão o controle desse conhecimento. São os diferentes campos de interesse da inteligência humana tornando-se ciências particulares, na interpretação de ter um campo delimitado de pesquisa. Paralelamente, o conhecimento científico passa a ser fundamentado em princípios evidentes e demonstrações, quer sobre raciocínios experimentais, ou ainda sobre a análise das sociedades e dos fatos humanos.

Além de possuírem em comum buscar respostas as inúmeras inquietações da nossa espécie, um aspecto partilhado dos diversos campos da ciência é que eles se abastecem

de dados. De acordo com Rampazzo (2013), o objeto das ciências são os dados próximos, imediatos, perceptíveis pelos sentidos ou por instrumentos, pois, sendo de ordem material e física, são, por isso, suscetíveis de experimentação.

Diante de informações numéricas sobre fenômenos e objetos diversos, muitas questões podem surgir diante de objetivos variados. Por exemplo, há alguma estrutura interessante nos dados de estudo? Há algum(uns) que destoa(m) dos demais? Alguns dos atributos estão correlacionados? Existem grupos diferentes com características em comum? Grupos semelhantes com alguma diferença? Pode-se obter o valor de um ou mais atributos a partir de suas medidas em dados coletados anteriormente? Estas são só algumas das inúmeras questões levantadas quando se tem o registro de dados.

Com uma infinidade de dados presentes no atual mundo globalizado, devidos aos avanços tecnológicos e à redução de custos nos sistemas de aquisição e armazenamento, grandes oportunidades para o desenvolvimento e aplicação de novos métodos se fazem possíveis, tanto no processo de mineração de dados quanto no reconhecimento de padrões. Entende-se por mineração de dados (*data mining*) a etapa onde estes são transformados em informações. Segundo Brandt (2014), além de projetar e realizar um experimento, uma tarefa importante na ciência experimental é a avaliação precisa e a exploração completa dos dados obtidos, que representam a quantificação dos fenômenos de interesse.

Acontece, porém, que os dados reais geralmente são de alta dimensão. De outra forma, são compostos de muitas variáveis dificultando ou impossibilitando os procedimentos relacionados à sua análise. Neste sentido, a redução da dimensionalidade é uma operação fundamental que busca reduzir o número de características de um conjunto de dados, porém procurando manter a essência das informações. Ela se baseia na aplicação de transformações aos dados mapeando-os em espaços de menor dimensão com máxima manutenção da topologia ou, em outras palavras, procurando manter as relações de vizinhança entre os dados. Conforme Yang et al. (2003), técnicas de redução de dimensionalidade (RD) objetivam reduzir o espaço de características preservando ao máximo as relações topológicas dos dados.

Outro ganho com a redução é a possibilidade da visualização de dados multidimensionais. A oportunidade de observar os dados mapeados em menor dimensão permite ao pesquisador gerenciar da melhor forma as decisões sobre a análise que se deseja, uma vez que, em tese, a redução captura as informações importantes do conjunto. Desta forma, é um recurso cada vez mais importante em mineração de dados, sobretudo na fase de percepção inicial da estrutura dos dados a serem trabalhados. A viabilidade do processo de visualização, contudo, depende fortemente de métodos eficientes de redução, principalmente, em conjuntos complexos de dados, mas também na própria estrutura dimensional intrínseca dos mesmos.

1.1 **Métodos de redução de dimensionalidade**

Existem vários métodos de redução de dimensionalidade. Uma distinção básica é se são mais apropriados para tratarem de estruturas lineares ou não-lineares e isso define a base dos chamados métodos lineares de redução de dimensionalidade ou métodos não-lineares de redução de dimensionalidade. Cada método visa preservar alguma propriedade específica de interesse no mapeamento. Em Lee e Verleysen (2007), os diferentes métodos de redução recebem uma de duas possíveis classificações: os que se baseiam na preservação da distância, mais restritivos, ou os que baseiam-se apenas na preservação da topologia. Ambos os tipos denotam mapeamentos contínuos, sendo que os primeiros são mais restritivos. Nesta seção, são listados os principais métodos, suas vantagens e desvantagens, contudo, sem a preocupação de buscar algum critério para relacioná-los.

1.1.1 Análise de Componentes Principais (*Principal Components Analysis* - PCA)

De acordo com Jolliffe (2002), a análise de componentes principais (PCA) é uma técnica linear de redução de dimensionalidade. Basicamente, ela possibilita identificar padrões em dados possivelmente expressos por alguma correlação intrínseca, mas não aparente. Como os padrões nos dados podem ser difíceis de serem encontrados nos dados de alta dimensão, onde não é possível observá-los graficamente, a PCA é uma valiosa ferramenta na análise dos dados, inclusive, de fácil aplicação.

Outra grande vantagem da PCA é que, uma vez encontrados padrões nos dados, a abordagem ainda permite considerar apenas alguns componentes essenciais, justificando seu nome. Para tanto, sua implementação passa pelo cálculo do espectro dos autovalores e autovetores da matriz de covariância dos dados e os autovetores associados aos maiores autovalores determinam a direção de projeção na qual estes dados apresentam maior variância. Com isso, se é capaz de reduzir a dimensionalidade dos dados sem muita perda de informação. Esta técnica é comumente utilizada como uma ferramenta exploratória de dados e na construção de modelos preditivos.

Matematicamente, a PCA é um procedimento que utiliza uma transformação ortogonal de vetores para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores linearmente sem correlação. O número de componentes principais é sempre menor ou igual ao número de variáveis originais. Ademais, possibilita encontrar um mapeamento linear entre um espaço de alta dimensão (d) e um subespaço de menor dimensão (m), $m < d$, que captura a maior parte da variabilidade dos dados. Os principais componentes são os autovetores dominantes da matriz de covariância dos dados. O anexo A traz o algoritmo em detalhes.

Apesar da análise de componentes principais ser de fácil implementação, muitos

dados do mundo real têm características não-lineares que um mapeamento da PCA não consegue captar. Uma ampliação da técnica, entretanto, nomeada *Kernel* PCA—KPCA (SCHöLKOPF; SMOLA; MüLLER, 1997) permite uma extensão da utilização da PCA tradicional, com o foco nas componentes principais que são relacionadas de modo não-linear às variáveis de entrada. Mais informações sobre o método podem ser obtidas em Jolliffe (2002).

1.1.2 Escalonamento Multidimensional

De acordo com Kruskal e Wish (1978), o escalonamento multidimensional (*Multidimensional Scaling* - MDS) refere-se a uma classe de técnicas que usam de proximidades entre qualquer tipo de objeto como entrada buscando encontrar um conjunto m-dimensional de pontos Z cuja distância entre eles seja intimamente consistente com um conjunto medido de similaridades. A técnica, primeiro apresentada por Torgerson (1952), pode ser dividida em diferentes categorias (BORG, 2007; COX; COX, 2008): MDS clássico, MDS métrico e MDS não-métrico. Em sua versão clássica, o objetivo do MDS é preservar as similaridades entre pontos de dados no espaço de imersão de acordo com o espaço de entrada.

Segundo Douglas Carroll e Arabie (1998), esta técnica compreende uma família de modelos geométricos para representação de dados em uma ou, mais frequentemente, duas ou mais dimensões e um conjunto correspondente de métodos para ajustar tais modelos aos dados reais. Uma definição mais restrita limitaria o termo a modelos de distância espacial para semelhanças, diferenças ou outros dados de proximidade. Nesse sentido, Hout, Papesh e Goldinger (2012) definem o MDS como um mapeamento que transmite, espacialmente, as relações entre os itens, em que os itens semelhantes estão localizados próximos um do outro e os itens diferentes estão localizados proporcionalmente mais distantes. A partir dessa imersão, pode-se inferir as dimensões subjacentes de um conjunto de dados examinando subjetivamente a organização do espaço.

O MDS métrico tem como objetivo mergulhar os dados originais $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, em um espaço de menor dimensão de forma que as distâncias entre pares de pontos, $d(\mathbf{x}_i, \mathbf{x}_j)$, sejam preservadas. Para tal, uma matriz D_X de distâncias iniciais é criada contendo as distâncias entre pontos no espaço original. A distância aqui, de modo simples, trata-se da euclidiana. Feito isto, mapeiam-se as distâncias entre os dados no espaço original para um outro espaço, de modo que, ao encontrar um novo conjunto de vetores $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m$, estas distâncias sejam preservadas. Assim, dados similares se mantêm próximos na nova representação. Ao fazer esse mapeamento a técnica também permite reduzir a dimensionalidade de um conjunto de dados.

Para realizar tal procedimento, o algoritmo minimiza uma função custo apropriada. De modo simples, uma função custo é uma medida de quão afastado o modelo está em

termos da capacidade de estimar a distância entre os dados no espaço onde eles são mapeados. Das diferentes funções disponíveis, uma conhecida como "*strain*"(deformação - ρ_s) é a mais utilizada. Neste caso, o uso do MDS por meio desta função é conhecida como MDS clássica, com

$$\rho_s(D_Z) = \|J^\intercal(D_X^2 - D_Z^2)J\|_F^2 \tag{1}$$

Aqui, de acordo com Ihler (2003), J é matriz dita centralizadora, de modo que $J = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$, onde I_n é a matriz identidade $n \times n$, $\mathbf{1}$ é um vetor de uns e $J^\intercal$ é a sua matriz transposta. A norma de *Frobenius* para uma matriz X, $\|X\|_F$, é definida como $\sqrt{\sum_{i=1}\sum_{j=1}|x_{ij}|^2}$.

Um aspecto interessante deste formalismo é que, minimizar a função custo nesse caso, corresponde a calcular os autovetores dominantes da matriz $-\frac{1}{2}J^\intercal D_X^2 J$. Ainda conforme Ihler (2003), os m principais autovetores capturam os maiores componentes de variação em $J^\intercal D_X^2 J$ e assim fornecem as coordenadas dos dados no novo espaço de características.

Por outro lado, ao usar distâncias euclidianas, o MDS clássico tem a desvantagem de considerar pequenas e grandes distâncias de igual modo. Desse modo, dependendo da natureza dos dados, a técnica pode não capturar com precisão a verdadeira topologia da variedade onde os dados residem (principalmente estruturas não-lineares) e assim não conseguir mapear os dados satisfatoriamente.

Alternativas para essas dificuldades também surgiram com a intenção de melhorar o algoritmo. Variações do método usam, por exemplo, um parâmetro que leva em conta o peso das distâncias entre os dados de entrada. O mapeamento não-linear de Sammon, por exemplo, um método similar ao MDS métrico (ver Lee e Verleysen (2007)), usa uma constante c na função custo que é inversamente proporcional à distância nos dados de entrada. Deste modo, a preservação de distâncias longas tem menos importância do que a preservação de distâncias curtas e, portanto, o uso deste fator de ponderação na função custo é justificado. Dessa forma, o método consegue lidar melhor com variedades não-lineares. Um panorama completo sobre este conjunto de técnicas pode ser encontrado em Ghojogh et al. (2020).

1.1.3 Imersão Localmente Linear

A imersão localmente linear (*Locally Linear Embedding* - LLE) é um método de redução de dimensionalidade não-linear proposto por Roweis (2000), Saul e Roweis (2003). O mapeamento procura preservar as relações de vizinhanças dos dados de entrada quando mapeados em um espaço de baixa dimensão. A princípio, a hipótese é que os dados

pertençam a uma variedade de baixa dimensão, localmente linear, e assim, cada ponto de dados pode ser mapeado usando as informações de seus q vizinhos com pesos apropriados. Tais pesos devem capturar a estrutura subjacente à variedade. De acordo com Saul e Roweis (2003), o algoritmo obteve seu nome devido à natureza das reconstruções, isto é, são locais, já que apenas os vizinhos contribuem para cada reconstrução, e linear, já que nas reconstruções são considerados subespaços lineares. Ainda, o resultado do LLE pode ser generalizado para novos locais no espaço de entrada fazendo que a técnica seja um mapeamento explícito entre os espaços de baixa e alta dimensão. Ou seja, pode-se calcular o valor de uma nova saída correspondente a uma nova entrada a partir do modelo treinado sem a necessidade de rodar novamente o algoritmo com o novo dado de entrada.

O LLE pode ser utilizado em problemas não-lineares de redução da dimensionalidade. Segundo Dias (2012), seu procedimento de otimização é mais simples de implementar e não envolve mínimo local, tendo um custo computacional favorável quando comparado a métodos puramente lineares, como o PCA e o MDS. Ainda, a implementação do método utiliza um problema de autovalores esparsos, diferente dos problemas de autovalores densos presentes na PCA e no MDS e, com isso, é mais eficiente que os outros algoritmos em termos de custos computacionais. O anexo B traz o algoritmo em detalhes.

1.1.4 Mapas de características isométricas

O *isomap* (*isometric mapping*) é um método não-linear de redução de dimensionalidade que, assim como o método principal exposto nesse trabalho—mapas de difusão—usa a distância em grafos como uma aproximação da distância geodésica. Pode-se entender a geodésica como uma linha reta no espaço curvo ou, nesta aplicação, a curva mais curta ao longo da estrutura geométrica definida pelos pontos de dados. O *isomap* procura mapear a distância entre os pontos de dados no espaço original de modo que coincida com a distância euclideana correspondente no espaço de imersão. O método foi proposto por Tenenbaum (2000) e pode ser visto como uma generalização do MDS. A diferença entre eles é que o *isomap* aproxima a distância geodésica na estrutura por grafos, enquanto que o MDS utiliza a distância euclidiana.

Acontece, porém, que nem sempre se tem conhecimento sobre a estrutura geométrica do conjunto de dados. Para contornar este fato, é feita uma aproximação da distância geodésica entre os dados assumindo que, em uma pequena vizinhança (determinada por vizinhos mais próximos ou pontos dentro de um raio especificado), a distância euclidiana é uma boa aproximação para a distância geodésica. É também o modo pelo qual se baseia a noção de similaridade no LLE. Assim, a distância geodésica é aproximada como a soma das distâncias euclidianas ao longo do caminho de conexão mais curto.

Uma desvantagem do algoritmo isomapiano é que, pelo menos na versão inicial,

a aproximação das distâncias geodésicas, como previamente descrita, não é robusta à perturbação por ruído. A escolha errada dos vizinhos mais próximos, por exemplo, com dados maculados por significante parcela de ruído, pode alterar muitas entradas na matriz de distância geodésica, que por sua vez podem levar a uma imersão em um espaço de baixa dimensão muito diferente (e incorreta). Como exemplo, se o número de vizinhos for muito pequeno, o grafo da vizinhança poderá se tornar pouco representativo para aproximar os caminhos geodésicos com precisão. Melhorias têm sido realizadas neste algoritmo para fazê-lo funcionar melhor em conjuntos de dados esparsos e ruidosos.

1.2 **Origem dos mapas de difusão**

De acordo com Wang (2012), a matemática dos mapas de difusão foi primeiramente estudada por Coifman e Lafon em Lafon (2004) e Coifman e Lafon (2006). No entanto, os *kernels* de difusão do tipo gaussiano eram amplamente usados em aprendizado de máquina e *clusters* de dados antes de 2004. O método chamado *Laplacian Eigenmaps* - LE (automapas do laplaciano), segundo Lee e Verleysen (2007), pode ser considerado o precursor dos mapas de difusão e foi introduzido por Belkin (2003) e Belkin e Niyogi (2003). Tal método pertence à família de técnicas de NLDR (*Non-Linear Dimension Reduction*) baseadas na decomposição espectral.

Ainda conforme Lee e Verleysen (2007), a técnica de automapas do laplaciano nasce procurando reparar algumas deficiências de outros métodos espectrais como *isomap* e LLE, desenvolvendo uma abordagem local para o problema de redução de dimensionalidade não-linear. Nesse sentido, o LE está intimamente relacionado ao LLE, embora enfrente o problema de forma diferente: em vez de reproduzir pequenos retalhos lineares com base em cada ponto de dados (como faz o LLE), o LE se baseia em conceitos teóricos de grafos, como o operador laplaciano em um grafo.

A base da técnica de automapas do laplaciano está na minimização restrita de distâncias locais, ou seja, distâncias entre pontos de dados vizinhos. Dessa forma, a técnica se baseia na hipótese de que o conjunto de dados $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ contém um número n suficientemente grande de pontos situados sobre (ou próximo a) uma variedade suave de tal modo que essa variedade possa ser representada com boa precisão por um grafo $\mathscr{G} = (V_n, \mathscr{E})$ não direcionado. Nesta representação, V_n é o conjunto dos vértices do grafo (com cada vértice v_i do grafo associado um dado $\mathbf{x}_i$), e $\mathscr{E}$ o conjunto das arestas desse grafo, onde uma aresta conecta os vértices v_i e v_j se os pontos de dados correspondentes forem adjacentes.

Seja $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ o conjunto dos dados mapeados. Para alcançar o mapea-

mento, a técnica se baseia na minimização da seguinte função custo:

$$f(Z) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2 \tag{2}$$

onde as entradas w_{ij} da matriz simétrica W exibem as relações de adjacência entre dados vizinhos $\mathbf{x}_i$ e $\mathbf{x}_j$ com $w_{ij} > 0$ se os pontos de dados são similares e $w_{ij} = 0$, caso contrário.

Atente que $f(Z)$ é a soma de termos não negativos na forma $w_{ij} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|^2$, portanto $f(Z) > 0$. Se w_{ij} é grande, *i.e.*, $\mathbf{x}_i$ e $\mathbf{x}_j$ são muito similares, então, para que a soma permaneça pequena—é desejado minimizá-la—o vetor $\mathbf{z}_i$ precisa estar próximo de $\mathbf{z}_j$. Caso contrário, se eles forem não similares, w_{ij} é pequeno, e a diferença entre $\mathbf{z}_i$ e $\mathbf{z}_j$ não prejudicará a minimização. Esse é o comportamento desejado para a imersão.

Para a escolha da noção de similaridade, Belkin e Niyogi (2003) recomendam o uso do *kernel* gaussiano. O motivo para usar a função gaussiana para o peso é devido ao seu decaimento característico: para pontos próximos, a semelhança é alta, enquanto, se estiverem afastados, a similaridade é próxima de zero.

$$w_{ij} = \exp\left(-\frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{T} \right) \tag{3}$$

Aqui, o parâmetro T pode ser interpretado como uma temperatura em um núcleo de calor envolvido em equações de difusão (LEE; VERLEYSEN, 2007). Tomar $w_{ij} = 1$, por exemplo, equivale a tomar $T = \infty$ no núcleo de calor.

Uma vez que W é simétrica, a função $f(Z)$ pode ser escrita na forma matricial como segue:

$$f(Z) = tr(ZLZ^\intercal) \tag{4}$$

Nesta equação, tr significa o traço da matriz e L é matriz laplaciana do grafo $\mathscr{G}$, definida como

$$L = G - W, \tag{5}$$

onde G é uma matriz diagonal com entradas $g_{ii} = \sum_{j=1}^{n} w_{ij}$. Já Z é uma matriz $m \times n$, onde o i-ésimo vértice do grafo $\mathscr{G}$ é movido para sua i-ésima coluna com o mapeamento e, com isso, cada $\mathbf{z}_i = (z_i(1), z_i(2), \ldots, z_i(m))^\intercal$ é a representação m-dimensional dos dados de entrada $\mathbf{x}_i$.

$$Z = \begin{bmatrix} | & & | \\ \mathbf{z}_1 & \cdots & \mathbf{z}_n \\ | & & | \end{bmatrix} \tag{6}$$

A prova para a igualdade em (4) encontra-se no anexo C. É possível observar que L é positiva semidefinida, uma vez que $tr(ZLZ^{\intercal}) > 0, \forall \mathbf{z}_i \neq 0$.

De acordo com Lee e Verleysen (2007), minimizar $f(Z)$ com respeito a Z se reduz a encontrar as soluções do problema de autovalor generalizado $\theta G\mathbf{f} = L\mathbf{f}$ procurando pelos autovetores de L associados aos seus menores autovalores. Como L é uma matriz simétrica e positiva semidefinida, todos os autovalores são reais e não negativos.

Acontece ainda que $\theta G\mathbf{f} = L\mathbf{f}$ possui uma solução trivial. De fato, para $\mathbf{f} = \mathbf{1}$, onde $\mathbf{1}$ é um vetor de uns, tem-se que $W\mathbf{1} = G\mathbf{1}$ e, assim, $L\mathbf{1} = \mathbf{0}$. Assim, $\theta = 0$ é o menor autovalor de L associado ao autovetor constante $\mathbf{f} = \mathbf{1}$.

Uma abordagem equivalente para obter a redução de dimensionalidade com a técnica de LE consiste na normalização da matriz laplaciana (BENGIO et al., 2006):

$$\mathscr{L} = G^{-1/2} L G^{-1/2} \tag{7}$$

e o encontro de seus autovetores:

$$\mathscr{L} = U \Lambda U^{\intercal} \tag{8}$$

Os autovetores associados aos m menores autovalores (exceto o primeiro, que é zero) formam uma imersão m-dimensional para o conjunto de dados. Os autovalores são os mesmos do problema de autovalor generalizado, e os autovetores são relacionados pela expressão $\mathbf{u} = G^{1/2}\mathbf{f}$.

Em sua tese, Belkin (2003) provou que o laplaciano de um grafo de pontos uniformementes amostrados (densidade uniforme) se aproxima do operador de Laplace-Beltrami sobre uma variedade, e que suas autofunções podem ser usadas para reduzir a dimensionalidade. Mais tarde, Lafon (2004) mostrou que este resultado não pode ser estendido a densidades não uniformes, e expandiu a análise descrevendo um algoritmo que lida com densidades mais gerais. Os automapas laplacianos e também os mapas de difusão, a serem discutidos no próximo capítulo, são motivados por esses resultados.

1.3 Mapas de difusão na organização de imagens digitais: um exemplo motivador

O algoritmo de mapas de difusão, em investigações de diferentes contextos, mostrou-se útil em diversas aplicações. Entre muitas outras, a técnica foi aplicada com sucesso em análise de documentos (ALLAH; GROSKY; ABOUTAJDINE, 2008), expressão genética (XU et al., 2010), reconhecimento de grupo audiovisual (KELLER et al., 2010), no estudo de processamento de sinais de voz (TALMON; COHEN; GANNOT, 2013), reconhecimento facial (BARKAN et al., 2013), citometria de massa resolvida no tempo de reprogramação

celular (ANGERER et al., 2015), dados de células-tronco hematopoéticas e progenitoras (HAGHVERDI; BUETTNER; THEIS, 2015), análise de perfis de madeira (MOURA NETO; SOUZA; MAGALHÃES, 2019), na busca de falhas relacionadas ao atrito em sistemas mecânicos ou tempos de produção na indústria (SHEVCHIK et al., 2021), etc.

Nesta seção, é apresentada uma aplicação da técnica de mapas de difusão em imagens digitais. O objetivo é mostrar a efetividade da técnica na eficiente redução de dimensionalidade em uma coleção de imagens de um brinquedo infantil cujo ângulo de rotação vertical do brinquedo representa, essencialmente, a única diferença das imagens no conjunto de dados (Fig. 1). Os mapas de difusão devem procurar pelas similaridades entre os píxeis que formam as imagens, buscando uma organização eficiente. Este exemplo foi inspirado em uma rotina semelhante presente em Talmon, Cohen e Gannot (2013).

1.3.1 Organização das imagens

O primeiro passo realizado foi a captura de imagens de um brinquedo infantil (carrinho de peças da marca Lego ®) em diferentes ângulos usando a câmera de um celular de modo fixo, como mostrado na figura 1. Cada imagem tem 400 píxeis (largura) e 300 píxeis (altura) e o ângulo de rotação do brinquedo é a única variável independente controlável no conjunto. Pequenas perturbações como diferença de iluminação e erro angular certamente existiram, não constituindo, no entanto, em variáveis controláveis.

Foi possível observar, a posteriori, que a qualidade das fotografias, tiradas sob diferentes ângulos é adequada para este estudo. Logo abaixo do objeto, uma folha branca de suporte contém as marcações da divisão de um ângulo raso em 15 partes iguais, conferindo assim, intervalos de 12° de rotação em torno de um eixo localizado na parte traseira do carrinho (onde há dois faróis vermelhos). É possível notar também, além da diferença entre o ângulo de rotação, que a iluminação ambiente afeta o formato da sombra.

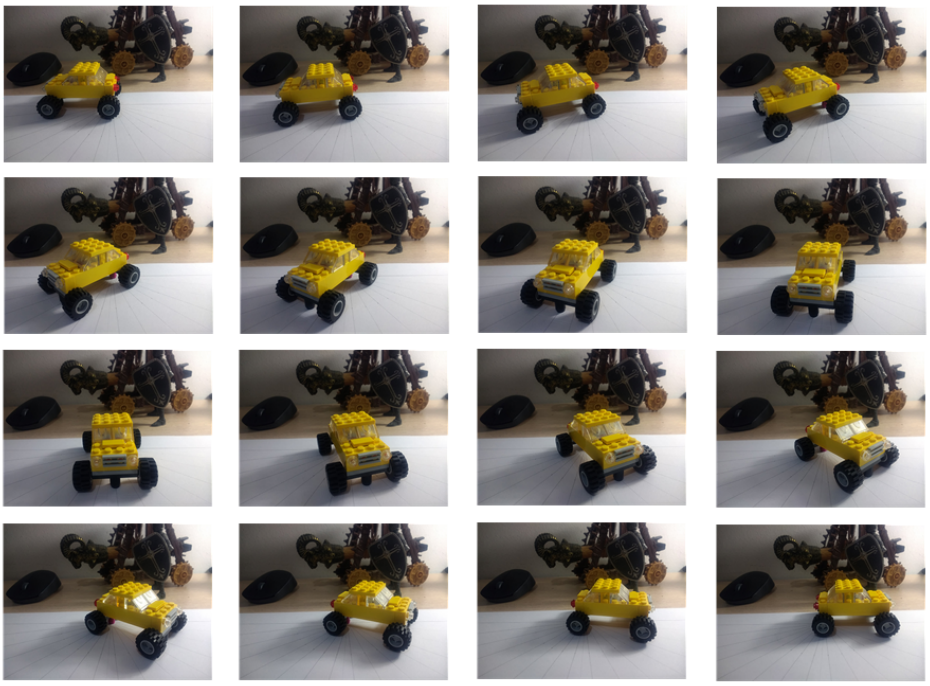
A tarefa aqui era a seguinte: dada a coleção de imagens em ordem aleatória como entrada, o algoritmo deve produzir o mapeamento eficiente dessas imagens e conseguir sua organização correta, produzindo como saída, as imagens na ordem crescente do ângulo de rotação.

Cada imagem foi inicialmente transformada em uma matriz de dados, e cada elemento é dado pela coordenada de um píxel no sistema de cores RGB. Desta forma, o algoritmo toma cada píxel como um vetor numérico de entrada e atribui como próximo, no sentido de vizinhança, o píxel que tem as coordenadas similares ao primeiro. A missão da técnica de mapas de difusão é, então, encontrar semelhanças entre as imagens representadas por vetores numéricos, a partir da noção de similaridade que esses dados pressupõem.

Após a execução do algoritmo e a obtenção do mapeamento em apenas uma dimensão (considerando somente o maior autovalor não constante), cada imagem é representada

como um número real e, para a organização eficiente das imagens, bastou colocá-los em ordem crescente. Dessa forma, a saída almejada, que independe da ordem de entrada das imagens, é mostrada na figura 2. Cabe informar que o procedimento foi repetido inúmeras vezes para diferentes configurações iniciais e, em todas elas, chegou-se à organização final correta.

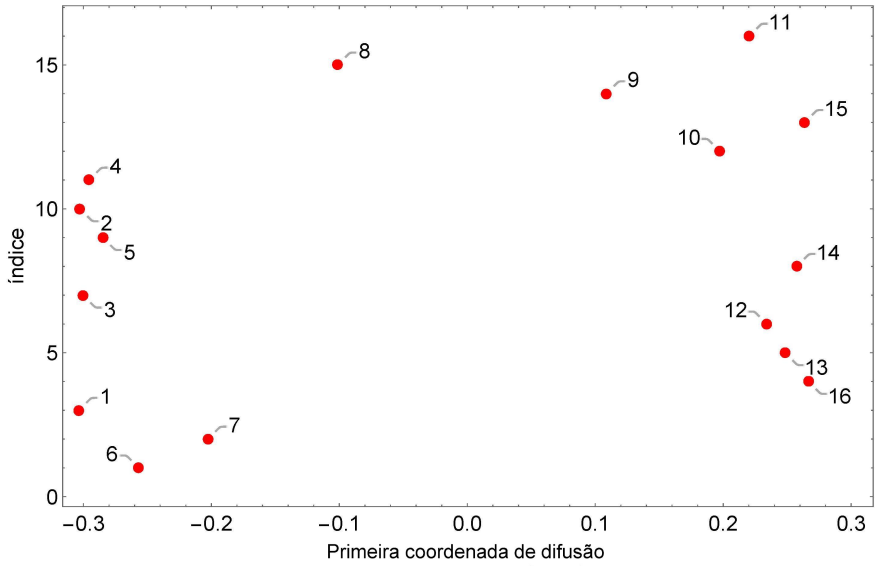
Figura 2 - Imagens do brinquedo infantil na ordem induzida pelas coordenadas do mapeamento 1D em ordem crescente, da esquerda para a direita, e de cima para baixo.



Fonte: O autor, 2022.

É interessante observar ainda a disposição dos dados mapeados pelos mapas de difusão 1D (Fig. 3). Na ordenada, tem-se o índice aleatório das imagens na entrada do algoritmo e, na horizontal, a primeira coordenada principal, resultado do mapeamento unidimensional. Os rótulos sobre cada ponto de dados mapeado exibem o índice correto das imagens ordenadas induzida pelos mapas de difusão. Verifica-se que a técnica mapeia os dados de modo a concentrar metade deles concentrados mais à esquerda, relativo às 8 primeiras imagens onde o ângulo de rotação está entre 0° e 90°, e as 8 seguintes com ângulo entre 90° e 180°. Isso evidencia que a técnica conseguiu absorver de modo suficiente o parâmetro que têm relevância na distinção das imagens (ângulo de rotação) e, com isso, conseguiu ordená-las de modo eficiente.

Figura 3 - Mapeamento 1D das imagens do brinquedo infantil.



Fonte: O autor, 2022.

1.4 **Ensaios de corrosão, curvas de polarização e perfis**

De acordo com Gentil (2011), a corrosão é definida como a deterioração de um material, geralmente metálico, por ação química ou eletroquímica no meio ambiente aliada ou não a esforços mecânicos. Nos ensaios de corrosão são caracterizados diversos aspectos da interação liga metálica-meio corrosivo, tais como o comportamento cinético das reações de oxirredução, a agressividade do ambiente corrosivo, e a morfologia da corrosão.

Basicamente, ensaios de corrosão podem ser feitos no laboratório ou no campo, dependendo dos objetivos que se quer alcançar. Gentil (2011) aponta que, enquanto nos ensaios de laboratório, consegue-se delimitar com maior controle as condições do experimento, buscando conseguir resultados mais rápidos, os ensaios de campo submetem o material em estudo diretamente às condições reais do meio corrosivo e os resultados são geralmente obtidos depois de longo período de tempo, sendo as condições de ataque muito diversas e às vezes não controláveis.

A fim de se avaliar quantitativamente o processo corrosivo e ter uma medida da extensão do ataque, vários métodos são empregados para verificar corpos-de-prova e, entre eles, pode-se citar a observação ao microscópio e os métodos eletroquímicos. Enquanto que a análise microscópica permite caracterizar ataques em nível microscópico como ataque transgranular, corrosão seletiva (como a dezincificação[1]), profundidade de pites e/ou

[1] A dezinficação é uma forma de corrosão seletiva que ocorre nas ligas de latão com mais de 15% de zinco. Consiste na corrosão do zinco, ficando a liga reduzida a um material esponjoso, constituído de cobre quase puro e sem resistência mecânica.

espessura de camada de revestimento, o ensaio eletroquímico é utilizado para medir a diferença de potencial entre diferentes metais, por meio de suas curvas de polarização em regiões catódicas e anódicas. Outros ensaios podem ser realizados, tais como caracterização físico-química e biológica do ambiente corrosivo, ensaios mecânicos, metalográficos e mesmo de modelagem computacional para se ter uma compreensão mais abrangente do processo corrosivo.

As curvas de polarização são uma importante caracterização nas investigações de uma variedade de fenômenos eletroquímicos. Segundo Stern e Geary (1957), tais medições permitem estudos do mecanismo de reação e da cinética dos fenômenos de corrosão e dissolução anódica do metal. A possibilidade de correlacionar o potencial aplicado e o logaritmo da densidade de corrente, permite observar regiões críticas onde, possivelmente, o processo de corrosão assume diferentes intensidades e comportamentos.

No campo da corrosão, assim como na ciência dos materiais, as curvas de polarização configuram importantes técnicas de pesquisa. Outros como diagramas de impedância, voltamogramas, ruído eletroquímico, entre outros, são de fundamental interesse para o estudo das complexidades inerentes à interação material - meio corrosivo.

Estes dados têm em comum o fato de serem expressos como perfis. Trata-se, então, de discretizações que permitem que estas curvas sejam imaginadas como vetores em $\mathbb{R}^n$ (n é o número de pontos da discretização) e, com isso, elas passam a ser vistas como dados de características não-lineares, ou seja, quando não há uma descrição de modelo simples. Neste sentido, é justificada a busca por técnicas de redução de dimensionalidade não-linear exploradas nesse trabalho—como ferramentas de análise e investigação na detecção de *outliers* e classificação/agrupamento dos resultados, levando em consideração a estocasticidade inerente aos dados experimentais.

2 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 estruturas importantes. A busca pela geometria intrínseca da variedade cujos dados em estudo se localizam deve permitir extrair informações relevantes e possibilitar sua compreensão. Para atingir qualquer um destes objetivos, tal método, assim como várias outras técnicas de mineração de dados e aprendizado de máquinas, empregam algoritmos baseados em grafos. Segundo Coifman e Lafon (2006), em termos de estruturas de dados, os grafos oferecem um compromisso vantajoso entre a simplicidade, a capacidade de interpretação e representação de relações complexas entre dados.

De modo sucinto, ao se utilizar grafos na representação do conjunto de dados e sua similaridade, o algoritmo permite definir uma métrica no espaço de distribuições de probabilidades que relaciona essas similaridades em múltiplas escalas utilizando-se de processos markovianos em pontos de dados que começam com similaridades locais de baixa ordem e evoluem rumo à integralização de toda a variedade. Ao final, o que se espera é um mapeamento que consiga capturar, de modo satisfatório, a similaridade entre os dados preservando a estrutura local.

Antes de descrever como aplicar a técnica e os passos de seu algoritmo, é necessário compreender intuitivamente como ela funciona e a teoria que sustenta todas as etapas do processo. Este capítulo busca discutir diferentes aspectos do algoritmo em detalhes.

2.1 Similaridades como propriedades das arestas de um grafo

Em diversas situações, no estudo de um fenômeno ou análise de diferentes processos nas mais variadas áreas do conhecimento, a coleta de dados é imprescindível. Como discutido no capítulo anterior, devido às diferentes naturezas dos registros, tais dados podem conter muitas características numéricas e, quase sempre, cada amostra pode ser representada por uma coleção de atributos numéricos cujo valor corresponde a uma característica observada ou mensurada (medida).

Nesse sentido, de posse de uma amostra de dados numéricos ou convertidos numericamente, o que se tem muitas vezes é um espaço de alta dimensão em que cada dado reside neste espaço e, a princípio, não se tem ideia da descrição global do conjunto. O conhecimento do espaço de características (*feature space*) dos dados, hipoteticamente, permitiria entender o processo/fenômeno que os gerou e viabilizaria conhecer sua interação. Na prática, isso possibilitaria a construção de modelos matemáticos eficientes para compreensão, predição e manipulação do fenômeno observado.

O primeiro passo na construção do algoritmo é medir a conectividade entre os dados observados. Nesse sentido, imagina-se cada ponto como o vértice de um grafo conectado cujas arestas ligam e têm seus pesos definidos de acordo com a similaridade relativamente aos seus vizinhos. Tem-se, então, um grafo com pesos que refletem uma similaridade local entre os dados.

Suponha agora que se faça um passeio aleatório nos pontos de dados, começando de um ponto específico, saltando-se aleatoriamente como um caminhante desnorteado, apenas guiado pela probabilidade de atingir cada ponto. O salto para um dado mais próximo é mais provável do que saltar para outro mais afastado. Nesta acepção, cria-se uma relação entre a distância no espaço de características 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 a 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*).

$$\text{conectividade}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) \tag{9}$$

Este núcleo define uma medida local de similaridade dentro de uma certa vizinhança. Fora dela, essa função precisa reduzir rapidamente a zero. Uma função que atende bem a este propósito é o conhecido núcleo gaussiano.

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\alpha}\right) \tag{10}$$

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 função $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon$ escolhido. Este parâmetro remete à conectividade do grafo e como o pesquisador imagina que os dados estejam relacionados. Na realidade, é o limite onde se imagina que a distância euclidiana entre os pontos de dados seja próximo da geodésica. 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 densidade dos dados. O último capítulo deste trabalho mostra como esta escolha é muito dependente dos dados e pode ser crucial diante do que se objetiva perante a aplicação requerida.

Sendo o núcleo de difusão um núcleo de similaridade, ele satisfaz as seguintes propriedades:

1. $\tilde{k}$ é simétrica, i.e. $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{k}(\mathbf{x}_j, \mathbf{x}_i), \forall i, j \in \mathbb{N}$.

2. $\tilde{k}$ preserva a positividade, i.e. $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \forall i, j$.

3. Dado um conjunto de dados $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, a correspondente matriz $\tilde{K}_{n\times n}$, onde $(\tilde{K})_{ij} = \tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$, é positiva semidefinida.[2]

Adiante é possível observar como as propriedades da função escolhida são necessárias para que o algoritmo funcione.

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

Obtém-se a matriz P utilizando-se a matriz de similaridade $\tilde{K}$,

$$(K)_{ij} = k_{ij} = \frac{\tilde{k}_{ij}}{\tilde{k}_i \tilde{k}_j}, \qquad \tilde{k}_i = \sum_{w=1}^{n} \tilde{k}_{iw} \tag{11}$$

$$(P)_{ij} = p_{ij} = \frac{k_{ij}}{y_i}, \qquad y_i = \sum_{w=1}^{n} k_{iw} \tag{12}$$

Define-se que a matriz Y, 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. Sendo assim, a matriz P como foi definida tem entradas não negativas e linhas que somam um 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 dessa cadeia.

De acordo com Coifman e Lafon (2006), do ponto de vista de análise de dados, a razão para estudar esta cadeia de *Markov* é que a matriz P contém informações geométricas sobre o conjunto de dados em estudo. Como foi definida, as transições expressas pela distribuição de difusão refletem a geometria local definida pelos vizinhos imediatos de cada nó no grafo de dados. 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 $\mathbf{x}_j$ sendo proporcional ao peso da aresta $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$. Tomando-se as potências da matriz de difusão P, aumenta-se o número de passos dados. À medida que a cadeia avança no tempo, o que equivale a tomar cada vez maiores potências de P, consegue-se revelar a geometria local e, com isso, ter conhecimento sobre as estruturas geométricas da variedade na qual o conjunto de dados está inserido.

Seja, por exemplo, uma matriz de difusão 3×3,

[2] Uma matriz M é positiva semidefinida se $\mathbf{x}^T M \mathbf{x} \geq 0, \forall \mathbf{x}$.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

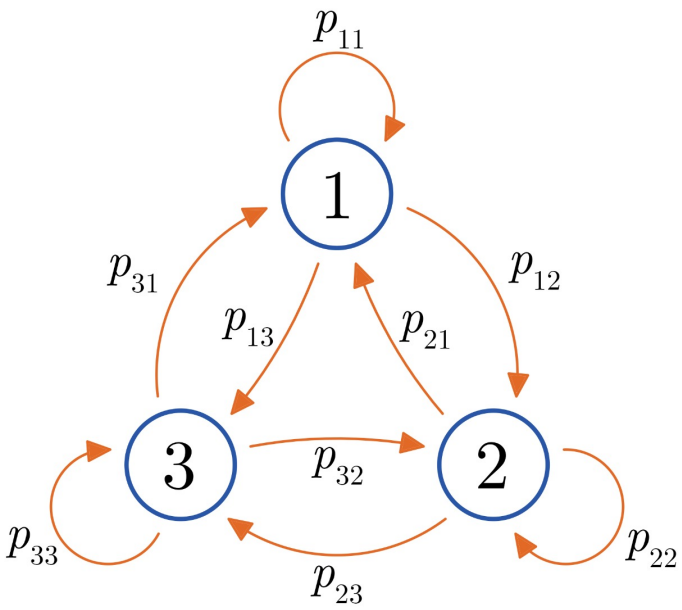
Cada elemento, p_{ij}, como viu-se, equivale à probabilidade de saltar entre pontos de dados i e j. Elevando P ao quadrado, tem-se:

$$P^2 = \begin{bmatrix} p_{11}p_{11} + p_{12}p_{21} + p_{13}p_{31} & p_{11}p_{12} + p_{12}p_{22} + p_{13}p_{32} & p_{11}p_{13} + p_{12}p_{23} + p_{13}p_{33} \\ p_{21}p_{11} + p_{22}p_{21} + p_{23}p_{31} & p_{21}p_{12} + p_{22}p_{22} + p_{23}p_{32} & p_{21}p_{13} + p_{22}p_{23} + p_{23}p_{33} \\ p_{31}p_{11} + p_{32}p_{21} + p_{33}p_{31} & p_{31}p_{12} + p_{32}p_{22} + p_{33}p_{32} & p_{31}p_{13} + p_{32}p_{23} + p_{33}p_{33} \end{bmatrix}$$

Ao pular dois saltos (isso, porque foi calculado a segunda potência), a matriz resultante mostra em cada entrada i, j a soma de todas as probabilidades do caminhante aleatório sair do estado i e alcançar o estado j em exatamente dois passos.

Observe que nesta matriz, por exemplo, o elemento da primeira linha e da primeira coluna soma três probabilidades: a de permanecer no estado 1 após ter permanecido no estado 1, a de voltar para o estado 1 depois de ter avançado para o estado 2 e a de voltar para o estado 1 depois de ter avançado para o 3. Da mesma forma, P^t_{ij} soma todos os caminhos de distância t do ponto i ao j. A figura 4 exibe um exemplo de grafo com 3 vértices e suas probabilidades de transição.

Figura 4 - Exemplo de um grafo com 3 vértices e suas probabilidades de transição.



Fonte: O autor, 2022.

À medida que t cresce, o processo de difusão prossegue com maior probabilidade de seguir um caminho ao longo da estrutura geométrica subjacente do conjunto de dados. Uma vez que os pontos de dados são bem concentrados sob a variedade, eles estão bem conectados. Com isso, os caminhos se formam ao longo de saltos curtos e de alta probabilidade. Em contrapartida, alguns caminhos são mais improváveis. No desenrolar do processo, a estrutura dos dados vai sendo explorada e a expectativa é que sua topologia seja gradualmente revelada.

2.2 **Distâncias de difusão e mapas de difusão**

Da seção anterior, concluiu-se que um processo de difusão utilizando as potências da matriz de *Markov* P é útil na descoberta da estrutura geométrica que contém os dados em várias escalas. Segundo Coifman e Lafon (2006), a cadeia de *Markov* define direções de propagação rápidas e lentas com base nos valores tomados pelo núcleo e, à medida que a caminhada avança, a informação da geometria local é propagada e acumulada podendo ser integrada para obter uma caracterização global do sistema.

Nesta seção, o objetivo é definir uma métrica com base nessa estrutura. Seja essa métrica a medida da similaridade de dois pontos de dados nesse espaço como a conectividade (probabilidade de "pular") entre eles. Como se nota, esta grandeza está relacionada com a matriz P e é dada por:

$$\begin{aligned} D_t(\mathbf{x}_i, \mathbf{x}_j)^2 &= \sum_{o \in X} \left| p_t(\mathbf{x}_j, o) - p_t(\mathbf{x}_i, o) \right|^2 \\ &= \sum_{e} \left| P^t_{je} - P^t_{ie} \right|^2 \end{aligned} \tag{13}$$

Observando a expressão, é possível notar que se trata da soma dos quadrados da diferença de cada elemento das linhas i e j da matriz P^t. A distância será pequena, como observado, se houver muitos caminhos de alta probabilidade com comprimento t entre dois pontos e permanecerá pequena enquanto as probabilidades de caminho entre $\mathbf{x}_i, o$ e $\mathbf{x}_j, o$ continuarem pequenas. Como a distância considera a soma entre todos os caminhos, a métrica é robusta a perturbações nos dados, ao contrário da aproximação isomapiana, por exemplo, para a distância geodésica.

A métrica definida anteriormente equivale à distância euclidiana entre os dados. A diferença é que esta métrica não é definida sobre suas coordenadas cartesianas, mas sim sobre suas coordenadas de difusão dadas pela matriz de difusão P. Assim sendo, espera-se que a distância detecte a similaridade de dois dados em termos dos seus parâmetros reais de mudança de acordo com a estrutura em que estão inseridos.

Se por um lado, a possibilidade do uso das distâncias de difusão é promissora, por

outro, já calcular essa distância para um conjunto pequeno de dados é computacionalmente penoso. A saída, então, é mapear os pontos de dados para um espaço euclidiano de acordo com a métrica definida. A distância de difusão no espaço de características torna-se a distância euclidiana neste novo espaço de difusão.

A seguir é mostrado como esse procedimento é realizado. Como o mapa de difusão deve preservar a geometria intrínseca do conjunto e o mapeamento deve absorver esta disposição em uma estrutura de mais baixa dimensão, espera-se descobrir que, de fato, são necessárias menos coordenadas para representar os pontos de dados no novo espaço. Essa hipótese é a chave para a redução de dimensionalidade. Contudo, quais dimensões devem ser rejeitadas? Como o mapeamento é realizado? Como calcular as distâncias de difusão otimizando o custo computacional? A seguir, tais questões são solucionadas.

Seja:

$$\tilde{\mathbf{z}}_i = \begin{bmatrix} p_t(\mathbf{x}_i, \mathbf{x}_1) \\ p_t(\mathbf{x}_i, \mathbf{x}_2) \\ \vdots \\ p_t(\mathbf{x}_i, \mathbf{x}_n) \end{bmatrix} \tag{14}$$

Definido desse modo, a distância euclidiana entre $\tilde{\mathbf{z}}_i$ e $\tilde{\mathbf{z}}_j$, é:

$$\begin{aligned} \|\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j\|^2 &= \sum_{o \in X} |p_t(\mathbf{x}_j, o) - p_t(\mathbf{x}_i, o)|^2 \\ &= \sum_{l} \left| P_{jl}^t - P_{il}^t \right|^2 \\ &= D_t(\mathbf{x}_i, \mathbf{x}_j)^2 \end{aligned}$$

que, como visto, é a distância de difusão entre os pontos de dados $\mathbf{x}_i$ e $\mathbf{x}_j$. Com este mapeamento, já é possível, por exemplo, a reorganização eficiente dos dados de acordo com a distância de difusão e a organização da coleção de imagens do brinquedo infantil (Fig.1) mostrada anteriormente. Contudo, nota-se que ainda nenhuma redução de dimensionalidade foi alcançada.

A redução de dimensionalidade é realizada negligenciando-se certas dimensões no espaço de difusão. Porém, antes de responder como isso é realizado, é necessário mostrar uma alternativa para os cálculos da distância de difusão.

Seja $\boldsymbol{\psi}_n$ os autovetores direitos de P associados aos autovalores λ_l, $l = 1, \ldots, n$. A distância de difusão entre $\mathbf{x}_i$ e $\mathbf{x}_j$ para um t fixo pode ser calculada como:

$$D_t^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{n} \lambda_l^{2t} (\psi_l(i) - \psi_l(j))^2 \tag{15}$$

Aqui, como apresentado, λ_l são os autovalores de P e $\psi_l(i)$ as coordenadas de seus

autovetores direitos.

$$P\boldsymbol{\psi}_l = \lambda_l \boldsymbol{\psi}_l \tag{16}$$

A prova da equação 15 encontra-se no anexo D.

A vantagem do cálculo da distância de difusão da maneira apresentada é que, como ela é dada em função dos autovalores e autovetores da matriz de *Markov* P, ela é menos custosa computacionalmente e de fácil implementação. A partir do cálculo da matriz P, o algoritmo calcula seus autovalores e autovetores e mapeia de acordo com as distâncias de difusão. Prontamente, o mapeamento pode, então, ser expresso em termos dos autovetores e autovalores de P como:

$$\mathbf{z}_i = \begin{bmatrix} \lambda_1^t \psi_1(i) \\ \lambda_2^t \psi_2(i) \\ \vdots \\ \lambda_n^t \psi_n(i) \end{bmatrix} \tag{17}$$

onde $\psi_1(i)$ indica o i-ésimo elemento do primeiro autovetor de P associado ao seu maior autovalor, $\psi_2(i)$ o i-ésimo elemento do segundo autovetor de P associado ao segundo maior autovalor, e assim por diante. Novamente, a distância euclideana entre os pontos mapeados $\mathbf{z}_i$ e $\mathbf{z}_j$ é a distância de difusão. Os autovetores ortogonais direitos de P formam uma base para o espaço de difusão e os autovalores, por sua vez, associados a esses, indicam a importância de cada dimensão. Enfim, a redução de dimensionalidade pode ser atingida escolhendo m dimensões associadas aos autovetores dominantes ($m < n$). Os elementos dos dados mapeados terão menos características numéricas do que os dados originais fazendo $\|\mathbf{z}_i - \mathbf{z}_j\|$ se aproximar da distância de difusão, $D_t(\mathbf{x}_i, \mathbf{x}_j)$.

O algoritmo básico da técnica de mapas de difusão é esquematizado a seguir.

Algoritmo

ENTRADA: Conjunto de dados de alta dimensão $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, parâmetro temporal t e a dimensão m pretendida para a redução.

1. Defina um núcleo $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$ e construa uma matriz $\tilde{K}$, de tal forma que $\tilde{K}_{ij} = \tilde{k}(\mathbf{x}_i, \mathbf{x}_j)$ represente a similaridade entre os pontos de dados $\mathbf{x}_i$ e $\mathbf{x}_j$.

2. Normalize as entradas da matriz de similaridade $\tilde{K}$, usando $(K)_{ij} = k_{ij} = \dfrac{\tilde{k}_{ij}}{\tilde{k}_i \tilde{k}_j}$, onde $\tilde{k}_i = \sum_{w=1}^{n} \tilde{k}_{iw}$.

3. Construa uma matriz de difusão P normalizando as linhas da matriz K através da relação $P = Y^{-1}K$, onde $(Y)_{ii} = \sum_{j=1}^{n} k_{ij}$.

4. Calcule os autovalores e autovetores da matriz P.

5. Mapeie os dados para o espaço de difusão m-dimensional, usando os m autovalores dominantes e seus autovetores de acordo com

$$\mathbf{z}_i = \begin{bmatrix} \lambda_2^t \psi_2(i) \\ \lambda_3^t \psi_3(i) \\ \vdots \\ \lambda_{m+1}^t \psi_{m+1}(i) \end{bmatrix}$$

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. $\lambda_2^t \psi_2(i)$ fornecerá as primeiras coordenadas de difusão do conjunto X mapeado para o novo espaço Z, $\lambda_3^t \psi_3(i)$ as segundas coordenadas de difusão, e assim segue.

SAÍDA: Conjunto de dados mergulhados em um espaço de menor dimensão $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^m$.

2.2.1 Mapas de difusão

Sejam $\psi_1, \psi_2, \ldots, \psi_n$ autovetores direitos de P associados, respectivamente, a $\lambda_1 = 1 \geq \lambda_2 \geq \ldots \lambda_n$. Para cada t fixo, chamam-se mapas de difusão as funções $\mathbf{\Psi}^{(t)}$, para cada $t = 0, 1, \ldots$, tal que:

$$\mathbf{\Psi}^{(t)}(\mathbf{x}_i) = (\lambda_2^t \psi_2(i), \lambda_3^t \psi_3(i), \ldots, \lambda_n^t \psi_n(i))^\intercal$$

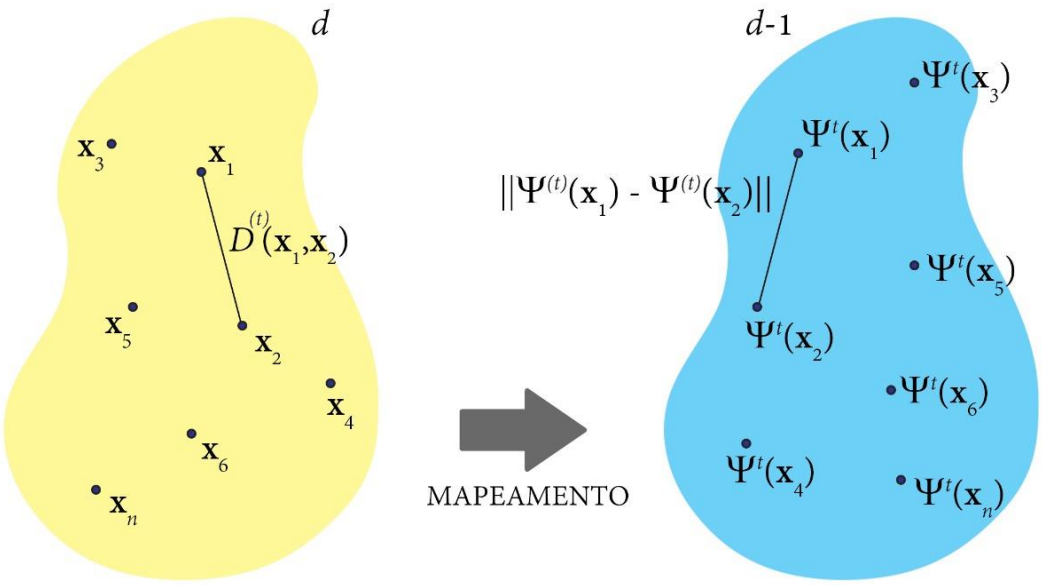
para cada $\mathbf{x}_i$ no conjunto X.

Com isso, a distância de difusão pode ser escrita em termos de mapas de difusão:

$$D^{(t)}(\mathbf{x}_i, \mathbf{x}_j) = \parallel \mathbf{\Psi}^{(t)}(\mathbf{x}_i) - \mathbf{\Psi}^{(t)}(\mathbf{x}_j) \parallel$$

A figura 5 exibe esquematicamente a relação entre as distâncias de difusão e os mapas de difusão.

Figura 5 - Relação entre as distâncias de difusão e os mapas de difusão.



Fonte: O autor, 2022.

3 MAPAS DE DIFUSÃO NO APRENDIZADO DE MÁQUINA DE SINAIS ELETROQUÍMICOS

Por meio da experimentação, informações quantitativas e qualitativas são extraídas da natureza pelas relações existentes entre os fenômenos observados, expressando as grandezas físicas por valores numéricos. Neste contexto, muitos dados oriundos de técnicas experimentais em ciência e engenharia de materiais estão disponíveis na forma de perfis (FABBRI et al., 2014), pois dependem fortemente de grandezas que variam sob uma faixa contínua, como a energia, comprimento de onda, temperatura, potencial, dentre outros.

Neste capítulo, faz-se a análise de perfis com dados originados de ensaios de corrosão. Trata-se de discretizações de curvas de polarização que passam a ser vistas como vetores. Estas curvas, por sua vez, de maneira mais geral, relacionam potencial e densidade de corrente e têm fundamental importância nos estudos de corrosão e de eletroquímica, sendo essenciais para medir a cinética global dos eletrodos.

O objetivo deste capítulo é a execução do método exposto anteriormente em dados reais experimentais expressos na forma de perfis unidimensionais provenientes de sinais eletroquímicos. Tais dados são os mesmos utilizados em Fabbri et al. (2014): dois aços inoxidáveis austeníticos comerciais (UNS S30400 e UNS S31600) expostos a meio aquoso com $3,5\%$ NaCl a $25°C$ e sob a taxa de polarização de $1,0$ mV/s para levantamento de curvas de polarização, aqui tratadas como perfis. No referido artigo, o objetivo foi basicamente avaliar a abordagem multi-q, usando a entropia de Tsallis, na classificação automática dos perfis avaliando o desempenho desta abordagem na separação eficiente dos perfis.

Na presente pesquisa, e especificamente neste capítulo, o propósito é semelhante - deseja-se atribuir um rótulo de classe (aço S30400 ou S31600) ao mesmo conjunto de dados e aplicar uma técnica de classificação para conseguir analisar os resultados para fins de discrimina-los. Nesta abordagem, por sua vez, utiliza-se dos mapas de difusão para o mapeamento e, para a classificação supervisionada, o mesmo conhecido *Naïve Bayes* ou, resumidamente, classificador *Bayes*. Mais informações sobre esse classificador podem ser obtidas em Witten (2011). O algoritmo foi implementado no ambiente *Wolfram Mathematica 12* (licença L4412-7008).

3.1 Curvas de polarização

Nesta seção é descrita a rotina experimental que deu origem aos dados utilizados neste capítulo, assim como em Fabbri et al. (2014). Dois aços inoxidáveis austeníticos comerciais foram expostos a um procedimento experimental na intenção de produzir suas

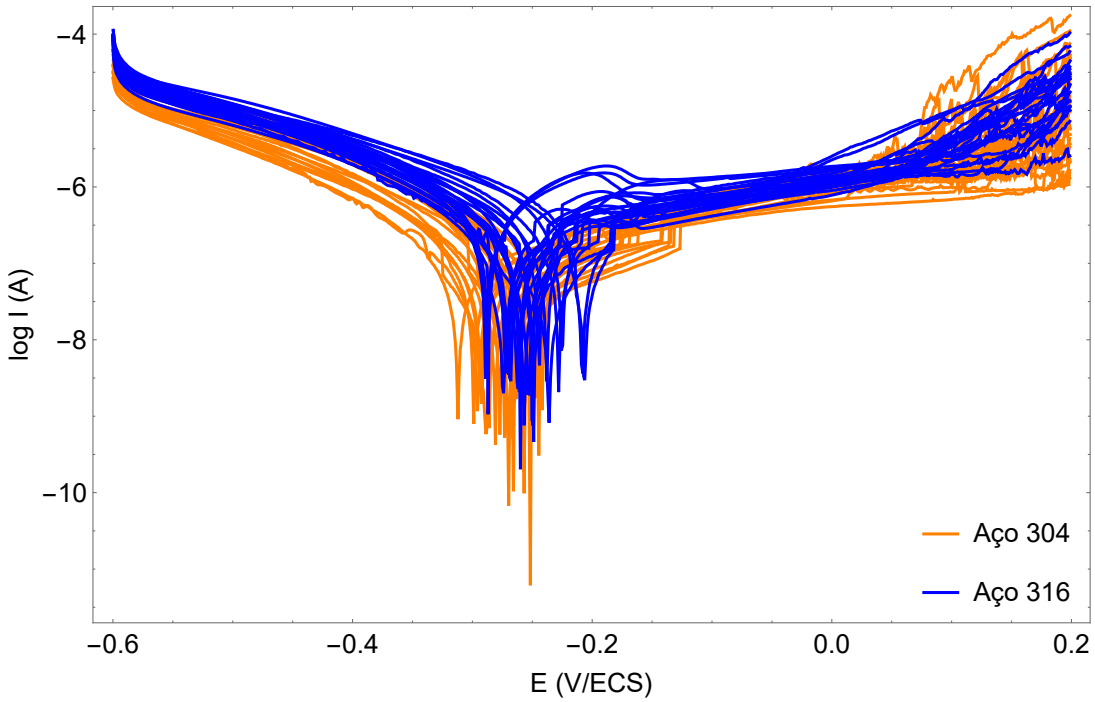
curvas de polarização. A excursão em potencial progride dos potenciais catódicos aos anódicos para dois aços inoxidáveis bem conhecidos.

O experimento utilizou dois aços inoxidáveis austeníticos comerciais (UNS S30400 e UNS S31600), a seguir denominados 304 e 316, por simplicidade. As amostras das barras foram revestidas com Teflon^{TM} e uma área de $0,20$ cm^2 foi mantida acessível ao eletrólito. Antes de cada ensaio, as amostras foram lixadas até lixa de granulometria #600, e posteriormente lavadas com água destilada e secas com ar quente. Além disso, após os testes de polarização, as superfícies foram analisadas com microscopia óptica para detectar a ocorrência de crévice. Todas as amostras apresentadas estavam livres de crévices após o ensaio de polarização. A principal diferença entre os aços é sua composição química, com o 316 contendo uma quantidade de cerca de $2,5\%$ de molibdênio; contra apenas vestígios encontrados no aço 304. Isso garante uma melhora na sua resistência à corrosão nos meios contendo cloreto principalmente em altos potenciais anódicos e fazem com que os aços ainda tenham resistência à corrosão localizada (YANG et al., 1984). Segundo Sun et al. (2021), o efeito do Mo reduz a densidade de corrente de corrosão e a expansão da corrosão localizada é dificultada, tornando o processo mais difícil de se propagar. O eletrólito aquoso é aerado com $3,5\%$ em massa de NaCl e a temperatura mantida em $25,0 \pm 0,2°C$.

Todos os experimentos foram realizados dentro de uma gaiola de Faraday para evitar ruídos elétricos espúrios e todos os valores do potencial foram medidos com uma referência de eletrodo de calomelano saturado (ECS). Após a imersão da amostra na solução, a condição desejada de estado estacionário é atingida após uma hora no status de circuito aberto. Após esta etapa, a polarização potenciodinâmica foi aplicada de $-0,6$ a $0,2$ V $\times$ ECS sob uma taxa de potencial de $1,0$ mVs^{-1}. Um fio de platina foi usado como contra-eletrodo. Para obter dados suficientes, foram obtidas 24 curvas para cada liga. Cada curva de polarização consiste em um perfil de corrente *versus* potencial obtidos com frequência de amostragem de um ponto por segundo, totalizando 800 pontos (Fig. 6). Nestas curvas de polarização, utilizou-se o valor absoluto da corrente (A) e não a densidade de corrente (A/cm²), porém o uso de densidade ou corrente não muda a análise realizada.

A figura 6 traz as 48 curvas de polarização dos aços 304 e 316 varridas das regiões de potencial catódico para anódico exatamente como é reportado em Fabbri et al. (2014). Como é possível observar, apesar de sua composição química diferir no teor de molibdênio, as curvas se sobrepõem em diversas faixas, principalmente, para os potenciais relacionados à predominância de processos anódicos ($-0,2$ a $0,2$ V $\times$ ECS). Nesta faixa 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. A probabilidade de classificar equivocadamente um aço em vez de outro é grande e é justamente na faixa de interesse do comportamento de corrosão desses tipos de aço. Isso justifica encontrar/combinar técnicas para melhor classificar os perfis de forma a obter resultados satisfatórios.

Figura 6 - 48 curvas de polarização experimental dos aços inoxidáveis 304 (laranja) e 316 (azul). Cada curva consiste em 800 pontos espaçados de 10 mV.



Fonte: O autor, 2022.

3.2 **Valores dos parâmetros de modelagem**

A escolha dos parâmetros de modelagem para a técnica de mapas de difusão é muito importante para a disposição final do mapeamento dos dados no espaço de difusão. Como observado nas equações 10 e 17, o método se utiliza de dois parâmetros nomeados α e t cujos valores devem ser escolhidos cuidadosamente diante da aplicação que se pretende tratar. Outra escolha ainda é referente à dimensão m do espaço que se objetiva fazer a redução de dimensionalidade, a qual será discutida mais adiante.

O parâmetro de escala α, por exemplo, pode ser escolhido de diferentes maneiras. Pode-se estabelecer uma faixa de valores que vai desde o limite de conectividade entre os pontos de dados, a menor das distâncias entre esses, até o diâmetro do conjunto, a maior distância. A depender dos dados, é comum também o uso de valores abaixo do limite de conectividade, o que resulta em um intervalo ainda maior. Este parâmetro reflete a escala considerada para definir a noção de similaridade entre os dados pertencentes a um espaço métrico definido.

Como visto, o parâmetro α tem influência direta no núcleo de similaridade e remete à conectividade do grafo e como se imagina que os dados estejam relacionados. Este parâmetro delimita os *clusters* (agrupamentos de dados afins em algum sentido) e desta forma, a escolha deste parâmetro deve ser selecionada. Acontece, porém, que, como

observado em alguns trabalhos, a justificativa para o uso do valor adotado desse parâmetro não é clara e, se algum critério é empregado, os trabalhos não o apresentam. Exemplos podem ser encontrados em Barkan et al. (2013), Salhov et al. (2015), Luo et al. (2015) ou Chen et al. (2016).

É visto neste capítulo que, pelo menos para os dados em estudos aqui abordados, a escolha do parâmetro que define a similaridade entre os pontos de dados é crucial e tem forte efeito nos resultados obtidos com a técnica. O capítulo 5, adiante, traz uma contribuição na análise minuciosa do efeito deste parâmetro em perfis simulados. Na literatura, como orientação geral, de acordo com la Porte et al. (2008), para estruturas intrincadas, não-lineares, de menor dimensão, uma vizinhança pequena é escolhida. Para dados esparsos, por outro lado, uma vizinhança maior é mais apropriada.

Por ora, a escolha do parâmetro de escala α pode assumir uma das escolhas seguintes. Seja $\varepsilon(i) = \min_{i \neq j} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel$. Algumas das escolhas razoáveis do parâmetro $\alpha = \varepsilon^2$ são dadas a seguir:

- O limite de conectividade do conjunto de dados,

$$\varepsilon = \varepsilon_{\mathrm{m}} = \min_i \varepsilon(i) \tag{18}$$

- O parâmetro de escala médio,

$$\varepsilon = \varepsilon_{\mathrm{mean}} = \frac{1}{n} \sum_i \varepsilon(i) \tag{19}$$

- O parâmetro de escala *min-max* do conjunto de dados,

$$\varepsilon = \varepsilon_{\mathrm{mM}} = \max_i \varepsilon(i) \tag{20}$$

- O diâmetro do conjunto de dados,

$$\varepsilon = \varepsilon_{\mathrm{d}} = \max\{\parallel \mathbf{x}_i - \mathbf{x}_j \parallel, i, j = 1, \ldots, n\} \tag{21}$$

A interpretação aqui é que $\varepsilon(i)$, sendo a menor distância entre um ponto qualquer $\mathbf{x}_j$ e um ponto fixo $\mathbf{x}_i$ distinto, cria um referencial para todas as outras distâncias.

É útil observar que a escolha do parâmetro de escala *min-max* do conjunto de dados, ou o parâmetro de escala médio representam opções intermediárias entre o limite de conectividade e o diâmetro do conjunto de dados. Enquanto que ε_{m} é a menor das distâncias entre dois pontos de dados quaisquer, ε_{d} é maior das distâncias observada. Por sua vez, $\varepsilon_{\mathrm{mean}}$ fornece a média entre as menores distâncias entre um ponto fixo e todos os outros e $\varepsilon_{\mathrm{mM}}$ indica a maior delas. Com isso, $\varepsilon_{\mathrm{m}} < \varepsilon_{\mathrm{mean}} < \varepsilon_{\mathrm{mM}} < \varepsilon_{\mathrm{d}}$.

Na seção seguinte, utilizando-se os dados das curvas de polarização, é mostrado que, além de empregar uma das escolhas apresentadas acima, é muito útil para a aplicação do método observar a disposição dos dados mapeados para diferentes valores desse parâmetro.

Na prática, pode-se definir um intervalo fechado entre dois limites escolhidos (como o ε_{m} e ε_{d}, por exemplo) e observar o mapeamento por uma visualização dinâmica tomando diferentes valores de α com intervalos de comprimento fixo dentro desse espaço. O objetivo desse procedimento é monitorar de modo dinâmico a distribuição dos dados mapeados neste novo espaço à medida que o parâmetro varia. Se $(\varepsilon_{\mathrm{m}})^2 = 10$ e $(\varepsilon_{\mathrm{d}})^2 = 1000$, por exemplo, poderíamos variar α de 1 em 1 e com isso a visualização dinâmica seria composta de 991 quadros. Cada um destes quadros exibiria os mapas de difusão dos dados em estudo para algum valor do parâmetro nesse intervalo.

De forma adicional, é proposta também a construção de um gráfico da função M que dá uma indicação da variação global dos mapas de difusão em função da mudança de α. Tal diagrama, pode exibir informações substanciais sobre a sensibilidade do parâmetro de escala diante dos dados em estudo a cada alteração da disposição dos dados mapeados. Esse comportamento será visto adiante e com mais detalhes posteriormente no capítulo 5 onde são utilizados perfis simulados.

Outro parâmetro de modelagem para a aplicação da técnica é o parâmetro temporal t. Ele está associado à matriz de difusão P e é responsável pelo processo de difusão. De acordo com a equação 17, t influencia diretamente as coordenadas dos dados mapeados para o espaço de difusão, uma vez que, representa o expoente dos autovalores na expressão que define essas coordenadas. É mostrado que na implementação da técnica para depuração dos dados, o parâmetro desempenha um importante papel na suavização do mapa de cores, fornecendo informações sobre *outliers* no conjunto. Esta abordagem também será vista adiante no decorrer da análise dos resultados propostos.

3.2.1 Uma medida da variação dos mapas de difusão

A fim de complementar a análise da sensibilidade do parâmetro de escala α que aparece no núcleo de similaridade, introduz-se uma função que exibe a variação global dos dados mapeados a cada variação desse parâmetro. Tal expressão, quando apresentada graficamente, pode exibir informações substanciais sobre a sensibilidade desse parâmetro diante dos dados em estudo a cada alteração. É mostrado agora como tal função, nominada função M (ou função de variação global), é definida.

Seja um intervalo $[a, b]$ onde α será escolhido. O limite inferior a pode ser o quadrado do limite de conectividade (Eq. 18), por exemplo, e b o quadrado do diâmetro do conjunto (Eq. 21). A depender dos dados, é comum também o uso de valores abaixo do limite de conectividade. Nos experimentos envolvendo dados simulados mostrados no

capítulo 5 foi utilizado um limite mínimo correspondente a, aproximadamente, 10% desse limite de conectividade.

Para α entre a e b, e cada i, seja a função,

$$\begin{aligned}&[a,b] \longrightarrow \mathbb{R}^m \\ &\quad \alpha \longmapsto \mathbf{z}_i(\alpha)\end{aligned} \tag{22}$$

Desta forma, se $m = 2$, por exemplo, cada $\mathbf{z}_i(\alpha)$ é da forma $(\lambda_2^t(\alpha)\psi_2(i,\alpha), \lambda_3^t(\alpha)\psi_3(i,\alpha))$ e representa o mapeamento (pelo mapa de difusão, com t fixo) do ponto de dado i ($\mathbf{x}_i$) para determinado α em duas dimensões. Dividindo o intervalo $[a,b]$ em intervalos de comprimento p, define-se:

$$\begin{aligned}&M : [a, b-p] \longrightarrow \mathbb{R} \\ &\qquad \alpha \longmapsto M(\alpha) = \left(\sum_{i=1}^{n} \|\Delta \mathbf{z}_i(\alpha)\|^2\right)^{\frac{1}{2}},\end{aligned} \tag{23}$$

onde $\Delta \mathbf{z}_i(\alpha) = \mathbf{z}_i(\alpha + p) - \mathbf{z}_i(\alpha)$. Assim sendo, $M(\alpha)$ dá a soma de todas as distâncias percorridas pelos dados mapeados a cada variação de α.

Pode-se pensar em escrever tal função deixando expostos os autovalores e autovetores que participam desta a cada variação do parâmetro. Substituindo a equação 17 na equação 23,

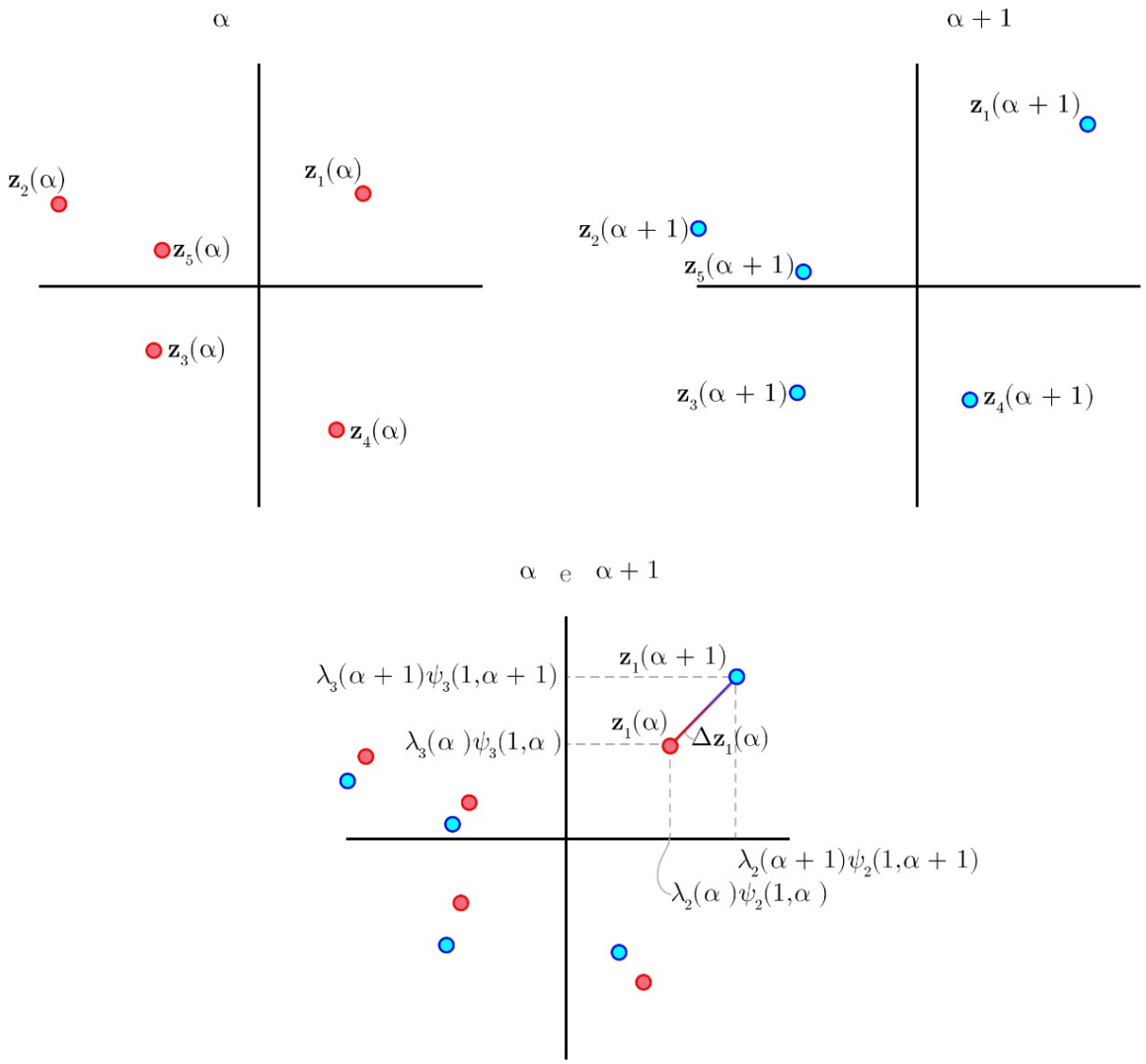
$$M(\alpha)^2 = \sum_{i=1}^{n} \|\Delta \mathbf{z}_i(\alpha)\|^2 = \sum_{i=1}^{n} \sum_{k=2}^{m+1} [\lambda_k^t(\alpha + p)\psi_k(i, \alpha + p) - \lambda_k^t(\alpha)\psi_k(i,\alpha)]^2. \tag{24}$$

Se $t = p \overset{N}{=} 1$ e, novamente, $m = 2$, por exemplo, o mapeamento obtido é bidimensional, o parâmetro α varia de 1 em 1 e a equação 24 toma a forma:

$$M(\alpha)^2 = \sum_{i=1}^{n} [\lambda_2(\alpha+1)\psi_2(i,\alpha+1) - \lambda_2(\alpha)\psi_2(i,\alpha)]^2 + [\lambda_3(\alpha+1)\psi_3(i,\alpha+1) - \lambda_3(\alpha)\psi_3(i,\alpha)]^2 \tag{25}$$

A figura 7 ilustra o caso para duas dimensões com $t = p \overset{N}{=} 1$.

Figura 7 - Esquema do cálculo de $M(\alpha)$ para duas coordenadas principais com $t = p \overset{N}{=} 1$. A função exibe a soma das distâncias percorridas por cada dado mapeado em relação a α.



Fonte: O autor, 2022.

Pode-se adotar também um ponto de vista contínuo e mais simplificado da função de variação global $M(\alpha)$. Suponha que o conjunto de dados, representado aqui por X, seja uma variedade diferenciável, $\mathbf{\Psi}$ seja o mapa de difusão para um t fixo e m o número de coordenadas de difusão. Assim,

$$\mathbf{\Psi} : [a, b] \times X \longrightarrow \mathbb{R}^m$$
$$(\alpha, \mathbf{x}) \longmapsto \mathbf{\Psi}(\alpha, \mathbf{x}) = (\lambda_2^t(\alpha)\psi_2(i, \alpha), \lambda_3^t(\alpha)\psi_3(i, \alpha), \ldots, \lambda_{m+1}^t(\alpha)\psi_{m+1}(i, \alpha)) \quad (26)$$

Uma forma de verificar a dependência de $\mathbf{\Psi}$ com respeito a α é considerar a derivada

parcial de $\mathbf{\Psi}$ com relação a α,

$$\frac{\partial \mathbf{\Psi}}{\partial \alpha} \in \mathbb{R}^m \tag{27}$$

Uma medida local dessa variação é então dada pela norma

$$\left\| \frac{\partial \mathbf{\Psi}}{\partial \alpha} \right\|. \tag{28}$$

Assim, pode-se definir uma medida global da variação de $\mathbf{\Psi}$ em função de α por

$$M(\alpha) = \left(\int_X \left\| \frac{\partial \mathbf{\Psi}}{\partial \alpha}^2 \right\| d\mathbf{x} \right)^{\frac{1}{2}}. \tag{29}$$

$M(\alpha)$ como definida anteriormente, é uma discretização desta quantidade.

3.3 **Abordagem de classificação**

Nesta seção é apresentada a metodologia para a classificação de padrões cujo objetivo básico é a atribuição de um rótulo de classe de aço a qualquer curva para a qual o tipo de aço é anteriormente desconhecido. Nessa tomada, a classificação é supervisionada, e o classificador é treinado a partir de um conjunto de amostras conhecidas—conjunto de treino—idealmente numeroso. A eficiência do classificador está na generalização de um conjunto de amostras—conjunto de teste—cujo rótulo é conhecido, mas que não tenham sido usadas para o treino. A rotina de classificação passa pelos processo de treino, validação e teste.

Nesta abordagem, os passos realizados foram os seguintes:

1. Aplicou-se a técnica de mapas de difusão ao conjunto de dados referentes aos perfis dos aços 304 e 316 usando o núcleo gaussiano e experimentando vários parâmetros de escala (α) e parâmetro temporal (t).

2. Obteve-se o mapeamento de acordo com o algoritmo reduzindo a dimensão de 800 para 3. O objetivo desta redução é a representação eficiente e também para fins de visualização. A redução para três dimensões configura o limite entre a possibilidade de visualização do mapeamento e maximização do número de informações consideradas.

3. Avaliou-se objetivamente o desempenho da abordagem de classificação, usando o esquema de validação cruzada de 10 vezes (WITTEN, 2011). Nesse modelo, o conjunto de dados do perfil é dividido aleatoriamente em 10 dobras (uma partição do conjunto de dados), considerando que cada dobra contém dois perfis, um para cada

classe do aço. Em cada execução deste esquema, o classificador é treinado usando todas as dobras menos uma dela e, em seguida, é avaliado como classifica as amostras da dobra separada. Gerou-se um número único para a taxa de acertos—uma medida de desempenho—com a classificação que representa a proporção geral de sucesso em todas as execuções.

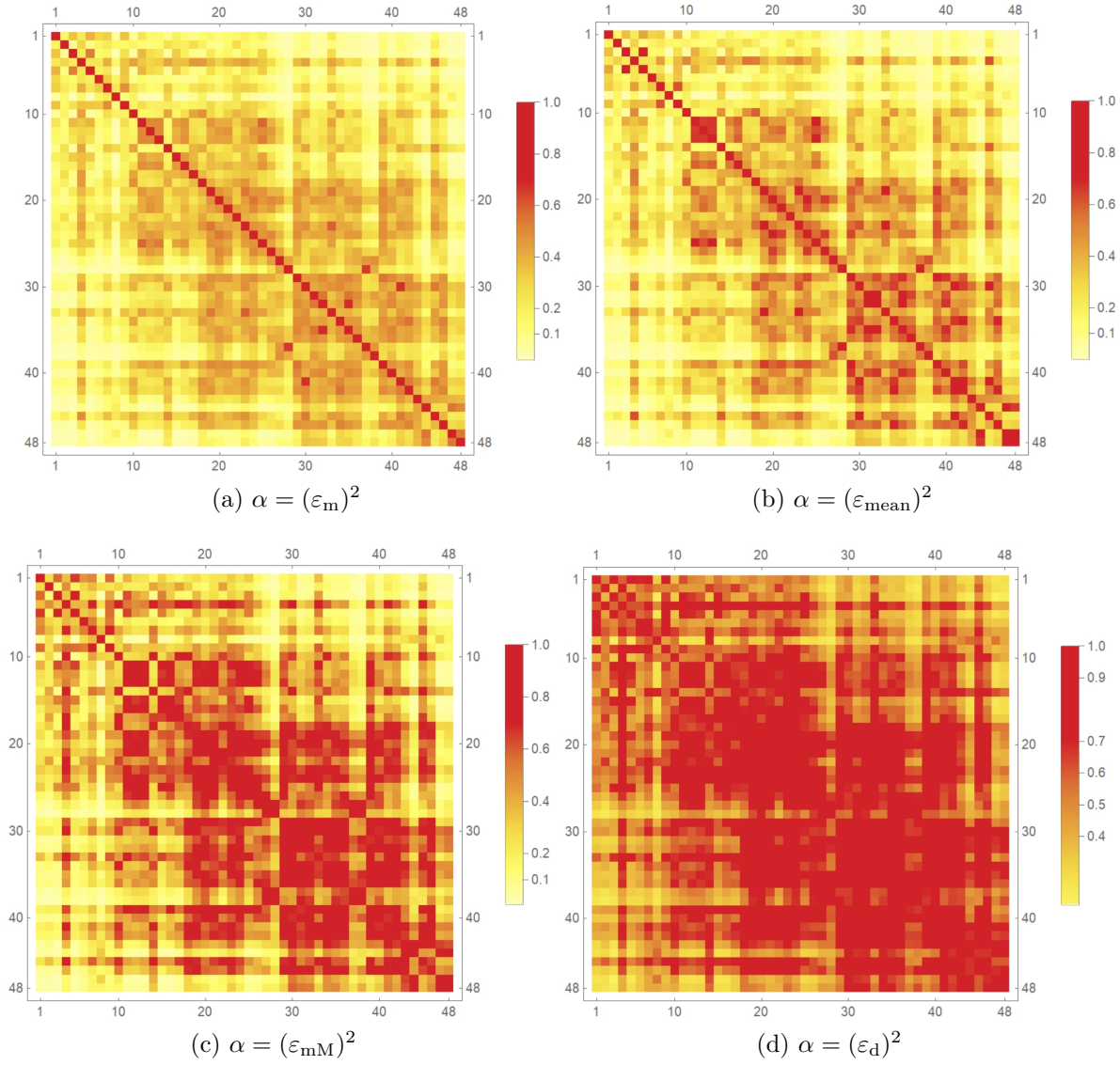
3.4 **Análise do parâmetro de escala do núcleo de difusão**

A primeira análise em questão dentro do que se propõe neste capítulo é o efeito do parâmetro de escala α no mapeamento do conjunto de dados. Na prática, como visto, isto reflete o tamanho da vizinhança considerada para a construção da matriz de similaridade. Um valor alto tornaria todos os dados bastante semelhantes entre si e, por outro lado, um valor baixo tornaria os dados dissimilares a quase todos os demais.

A figura 8 traz representações das matrizes de similaridade para os diferentes parâmetros de escala α apresentados na seção anterior. A matriz $\tilde{K}$ é vista como uma função $(i, j) \mapsto (\tilde{K})_{ij} \in \mathbb{R}$ e as subfiguras representam os níveis dessa função utilizando cores para representar os valores de $\tilde{K}_{ij}$ para os diferentes parâmetros. Se $\tilde{K}_{ij}$ está próximo de zero, a cor é branca e, quando $\tilde{K}_{ij}$ está próximo de um, seu maior valor, a cor é vermelha.

É possível notar que com crescentes valores de ε (e de α, consequentemente), mais pontos (perfis) ficam próximos a outros, uma vez que, a figura 8(d) tem mais regiões avermelhadas que as figuras 8(c), 8(b) e 8(a). De fato, quando se permite que o parâmetro de escala cresça, a vizinhança dos dados se torna maior—os pontos de dados se tornam semelhantes a um número maior de outros—e isso tem um impacto na matriz de similaridade. Naturalmente, a diagonal principal é sempre bem vermelha uma vez que se trata da semelhança de um perfil com ele mesmo, representando a máxima similaridade possível. A figura 9 mostra o impacto do parâmetro α também nos autovalores da matriz de difusão P. Observa-se que o maior autovalor, independe de α, se iguala a 1. Conforme α cresce, mais rapidamente os autovalores decrescem.

Figura 8 - Representações das matrizes $\tilde{K}$ para diferentes escolhas do parâmetro de escala.

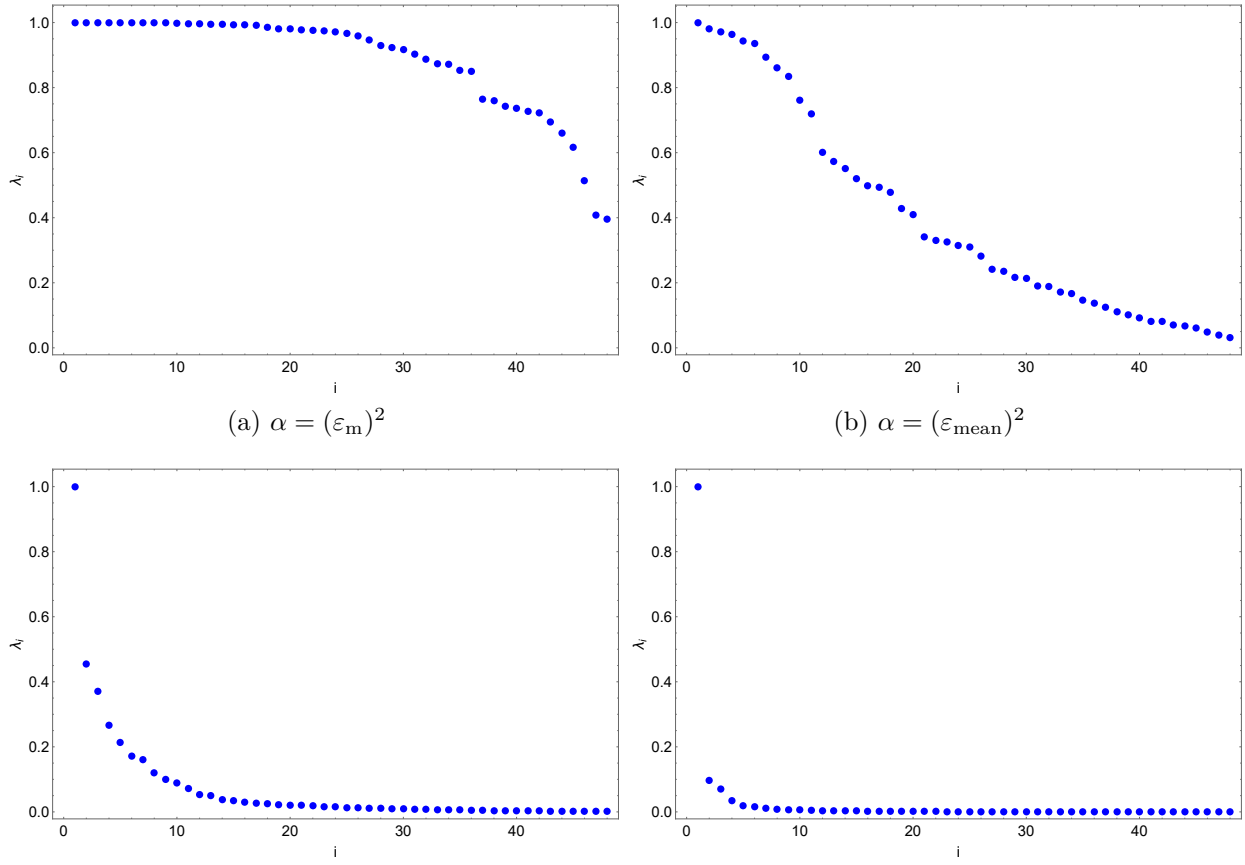


(a) $\alpha = (\varepsilon_{\mathrm{m}})^2$ (b) $\alpha = (\varepsilon_{\mathrm{mean}})^2$

(c) $\alpha = (\varepsilon_{\mathrm{mM}})^2$ (d) $\alpha = (\varepsilon_{\mathrm{d}})^2$

Fonte: O autor, 2022

Figura 9 - Autovalores da matriz de difusão P para diferentes valores do parâmetro α.



(a) $\alpha = (\varepsilon_{\mathrm{m}})^2$ (b) $\alpha = (\varepsilon_{\mathrm{mean}})^2$

(c) $\alpha = (\varepsilon_{\mathrm{mM}})^2$ (d) $\alpha = (\varepsilon_{\mathrm{d}})^2$

Fonte: O autor, 2022

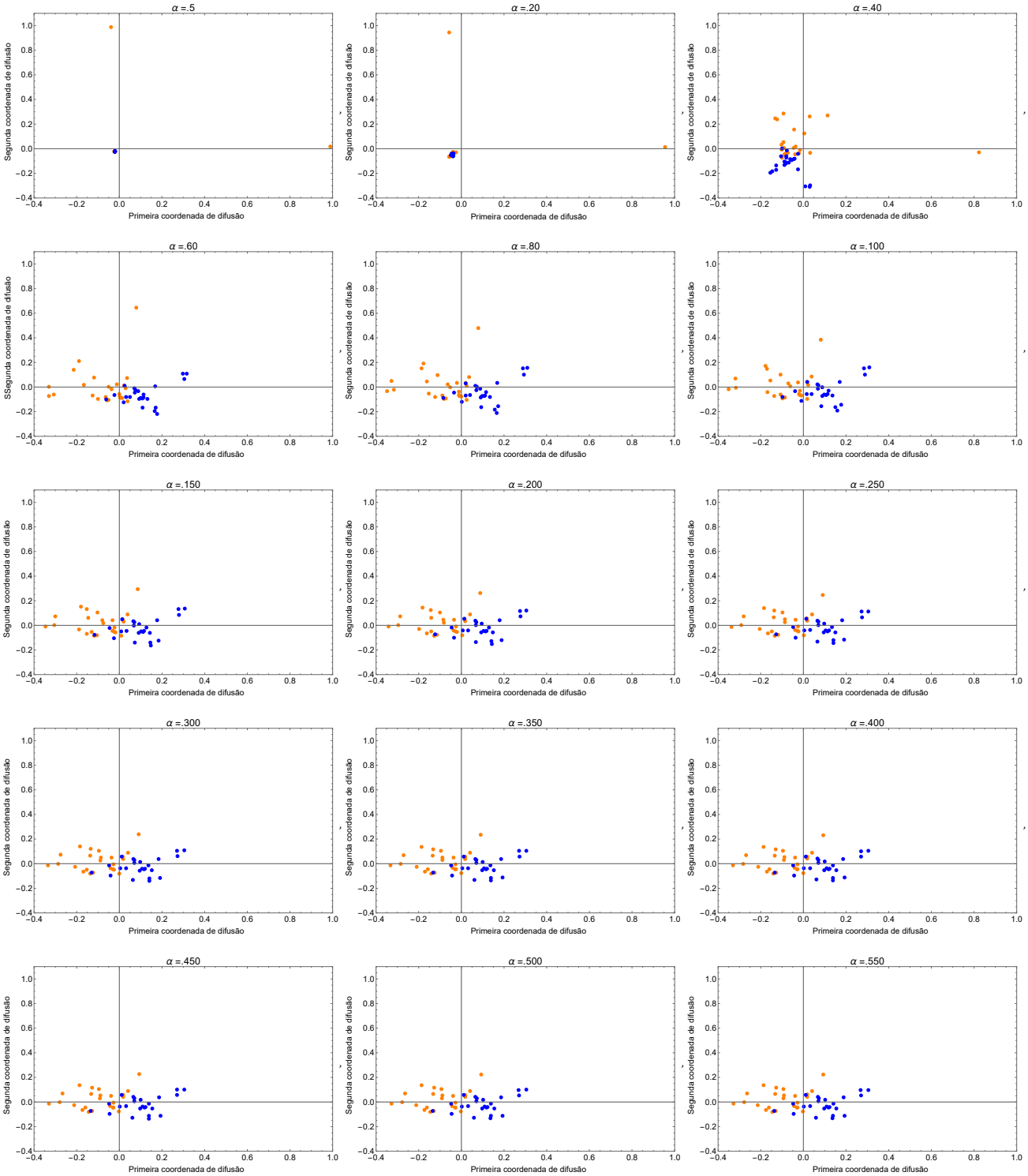
No processo de implementação, durante a experimentação e análise das saídas para cada valor do parâmetro α, observou-se alteração na disposição dos dados mapeados dependendo da escolha deste parâmetro. Para tal, fixou-se $t = 2$ e um intervalo fechado foi tomado com $5 \leq \alpha \leq 550$, observando o mapeamento obtido por visualização dinâmica tomando diferentes escolhas para α em intervalos de 1 em 1 dentro desta faixa. Cabe informar que os valores $\alpha = (\varepsilon_{\mathrm{m}})^2 = 7,989; \alpha = (\varepsilon_{\mathrm{mean}})^2 = 25,7467; \alpha = (\varepsilon_{\mathrm{mM}})^2 = 115,085$ e $\alpha = (\varepsilon_{\mathrm{d}})^2 = 546,691$ estão incluídos no intervalo de análise. A visualização dinâmica aqui se refere à análise de uma sequência de quadros que compõem um vídeo, que é o mapeamento 2D dos dados para um α específico. Em outras palavras, adotando α como um parâmetro temporal na evolução do vídeo. Com isso, pôde-se observar em detalhes o efeito do parâmetro no mapeamento. O vídeo encontra-se disponível no link <https://youtu.be/68dTMbajWac>. A figura 10 mostra alguns dos quadros.

Analisando-se as figuras 8 a 10, é possível notar a importância da escolha desse parâmetro. Primeiramente, como já comentado, a figura 8 exibe que essa escolha reflete no grau de similaridade dos dados. Em seguida, a figura 9 mostra que ela influencia também os autovalores da matriz de difusão e, assim, consequentemente as coordenadas dos dados mapeados. Um valor pequeno para este parâmetro de escala, como $\alpha = (\varepsilon_{\mathrm{m}})^2$, por exemplo, faz com que vários autovalores sejam próximos de um e, com isso, de acordo com a equação 17, essa escolha torna limitada o efeito do parâmetro t sobre as principais componentes do mapeamento, uma vez que se $\lambda_i \approx 1$, então $(\lambda_i)^t \approx 1, \forall t \in \mathbb{N}$.

Por outro lado, a escolha de um valor grande para α, como $\alpha = (\varepsilon_{\mathrm{d}})^2$, por exemplo, faz o decaimento dos autovalores ser muito acentuado e, como será considerado adiante, também pode não ser uma escolha adequada.

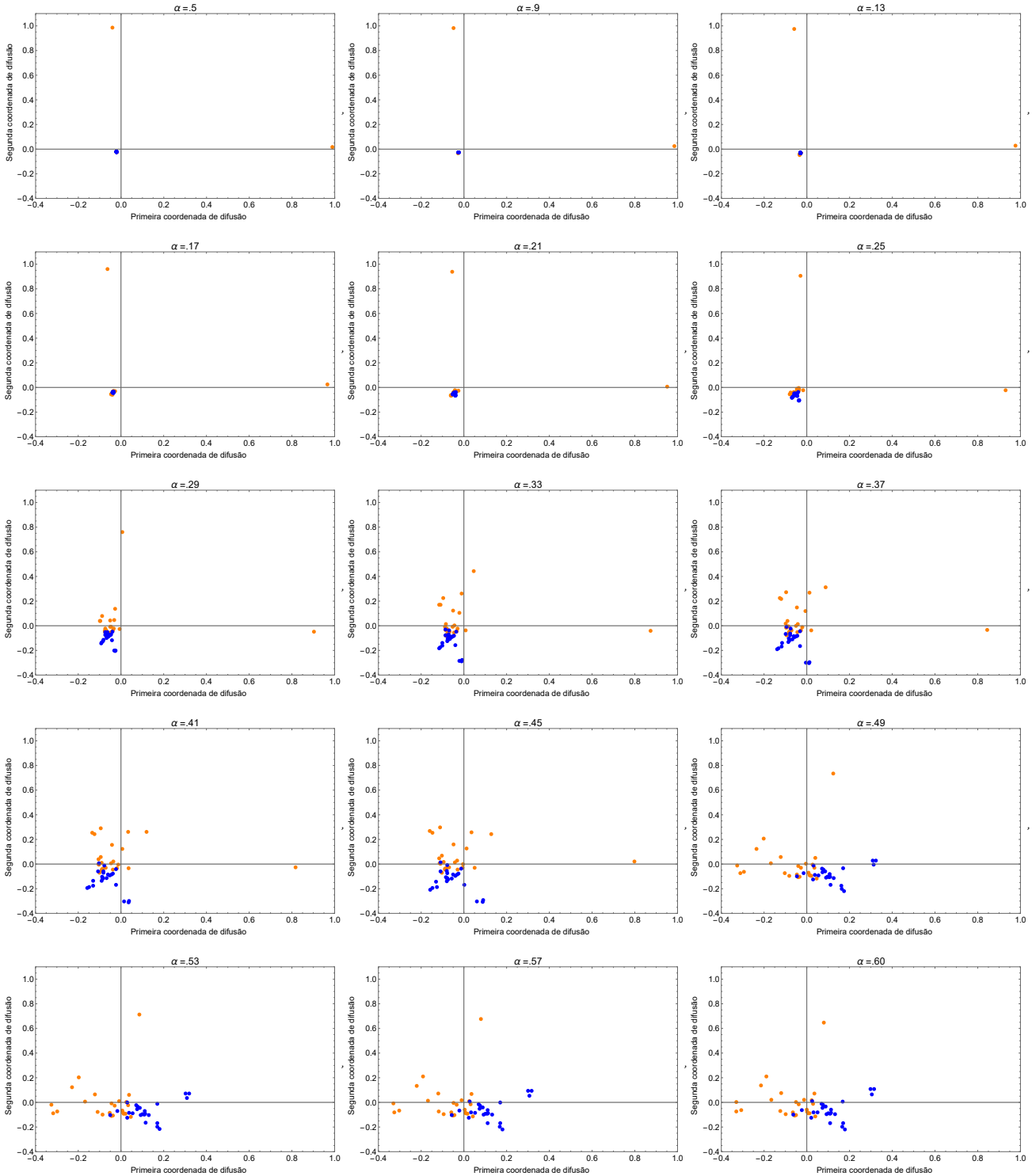
A figura 10, por sua vez, revela adicionalmente que o parâmetro α, além de ter influência no mapeamento, provoca, em determinados valores, mudanças abruptas na disposição dos dados mapeados, com súbitas transições qualitativas. Observa-se que a principal diferença entre os mapeamentos obtidos, em relação à disposição dos dados mapeados, ocorre antes de $\alpha = 60$. Neste trecho tem-se maior variação nas coordenadas destes dados evidenciado pela sua distribuição em função do valor do parâmetro. A figura 11 apresenta com detalhes os mapeamentos para essa faixa.

Figura 10 - Mapeamento 2D para diferentes valores do parâmetro α



Fonte: O autor, 2022.

Figura 11 - Mapeamento 2D para diferentes valores do parâmetro $\alpha \leq 60$.

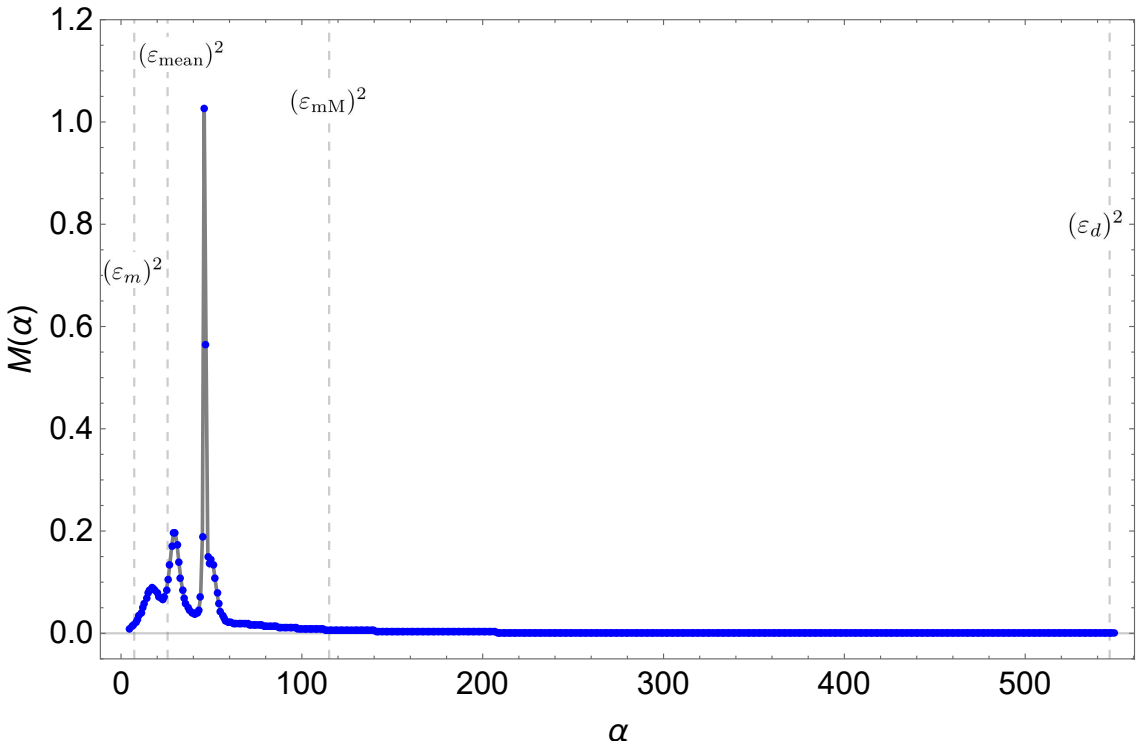


Fonte: O autor, 2022.

Como apresentado, obteve-se o mapeamento dos perfis começando de $\alpha = 5$ e seguindo passos de 1 em 1 até 550. A análise da figura 12 corrobora o que foi descrito anteriormente acerca da sensibilidade dos mapas de difusão com respeito ao parâmetro de escala α no intervalo $5 < \alpha < 60$. Nesta região, a disposição dos dados mapeados se altera mais ativamente a cada mudança do parâmetro evidenciado pelos picos presentes no gráfico.

Pode-se elencar duas observações interessantes destes resultados. A primeira delas é que a disposição dos dados mapeados não mais varia a partir de $\alpha = 60$, aproximadamente. Isto pode ser observado pelo comportamento decrescente e assintótico da curva acima deste valor. Assim, se o interesse na utilização da técnica é a separação eficiente dos perfis, que é um dos objetivos do capítulo, esse mapeamento não mais traz vantagens para a partição dos *clusters* de dados que representam os aços mapeados, com o efeito do parâmetro alfa. A segunda conclusão é que a presença de picos e vale nesse trecho de sensibilidade da curva parece mostrar, justamente, uma mudança de percepção da técnica em relação à estrutura dos dados, que ora os enxerga como um só grupo e ora como dois, dependendo da escolha do parâmetro.

Figura 12 - $M(\alpha)$ para $5 \leq \alpha \leq 550$. O parâmetro de escala é mais sensível para valores entre o quadrado do limite de conectividade e o quadrado do parâmetro *min-max* do conjunto de dados.

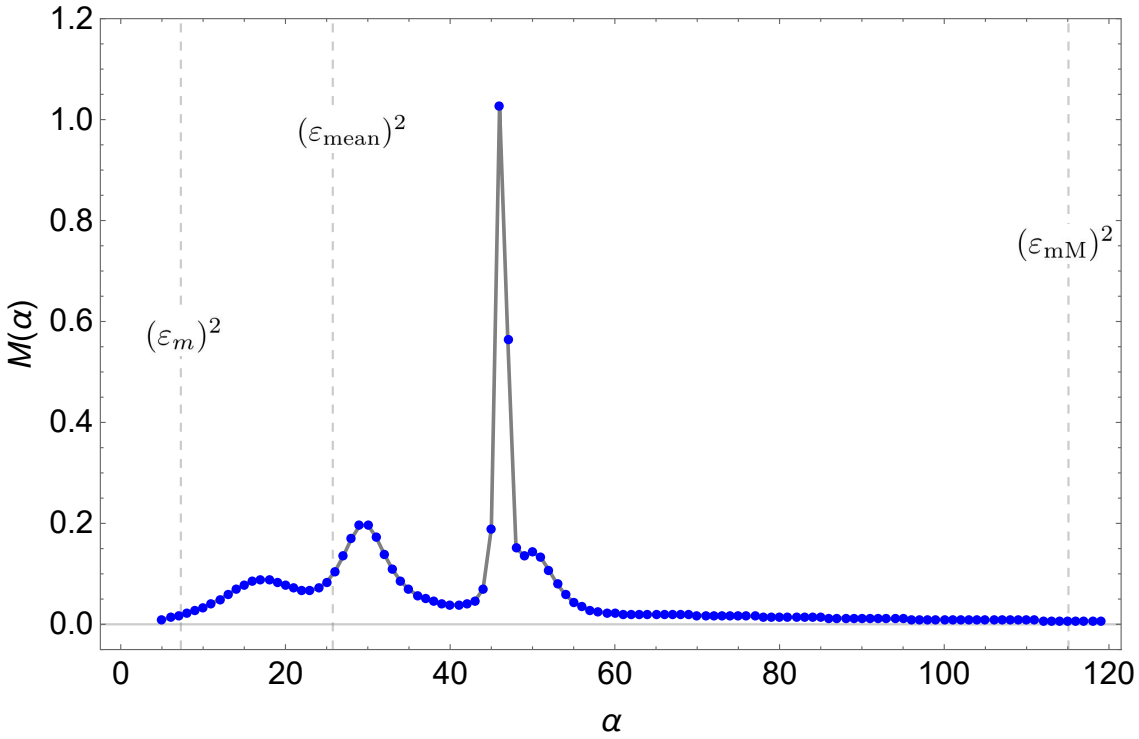


Fonte: O autor, 2022.

A figura 13 mostra em mais detalhes essa região de sensibilidade. Nesta imagem,

obteve-se $M(\alpha)$ para o mapeamento dos perfis começando de $\alpha = 5$ e seguindo passos de 1 em 1 até 120. Dessa forma, consegue-se examinar de modo acurado a mudança qualitativa na disposição dos dados para a região $5 \leq \alpha \leq 120$.

Figura 13 - $M(\alpha)$ para $5 \leq \alpha \leq 120$.

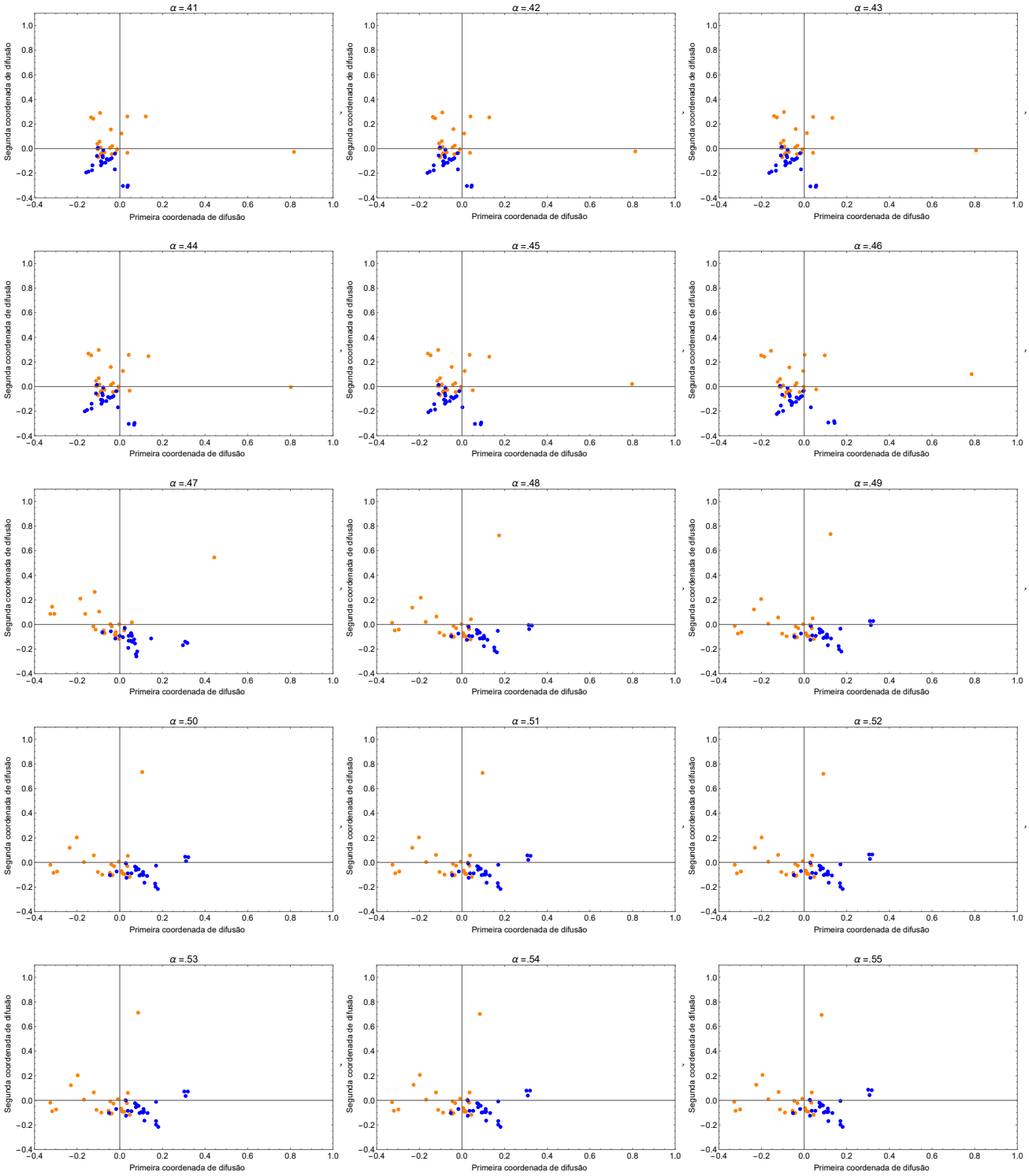


Fonte: O autor, 2022.

Fazendo-se a análise da figura 13, pelo menos três aspectos chamam novamente a atenção. O primeiro é a existência de mínimos locais da função M, que denotam possibilidades justificáveis para a escolha de α uma vez que representam situações de estabilidade do mapeamento: valores próximos não alteram significativamente a disposição dos dados mapeados. O segundo, são os pontos de máximo local da função, em que o mapeamento sofre uma grande alteração. Destaca-se, em especial, o ponto de máximo global contido na região de alta sensibilidade com entre $40 < \alpha < 55$, aproximadamente, que é onde acontece uma rotação na configuração dos dados mapeados (Fig 14). O terceiro aspecto é sobre a dificuldade de interpretação do comportamento sobre o que acontece com $\alpha < 5$.

Em relação à região de alta sensibilidade com $40 < \alpha < 55$ trata-se de uma região limite onde a percepção da técnica sobre a estrutura dos dados se modifica fortemente. Em outras palavras, a hipótese é que nesta região a técnica de mapas de difusão interpreta que há dois grupos de *clusters*, ainda que pouco distintos, o que justificaria a disposição dos dados mapeados não sofrer mais alteração significativa a partir dessa região de transição.

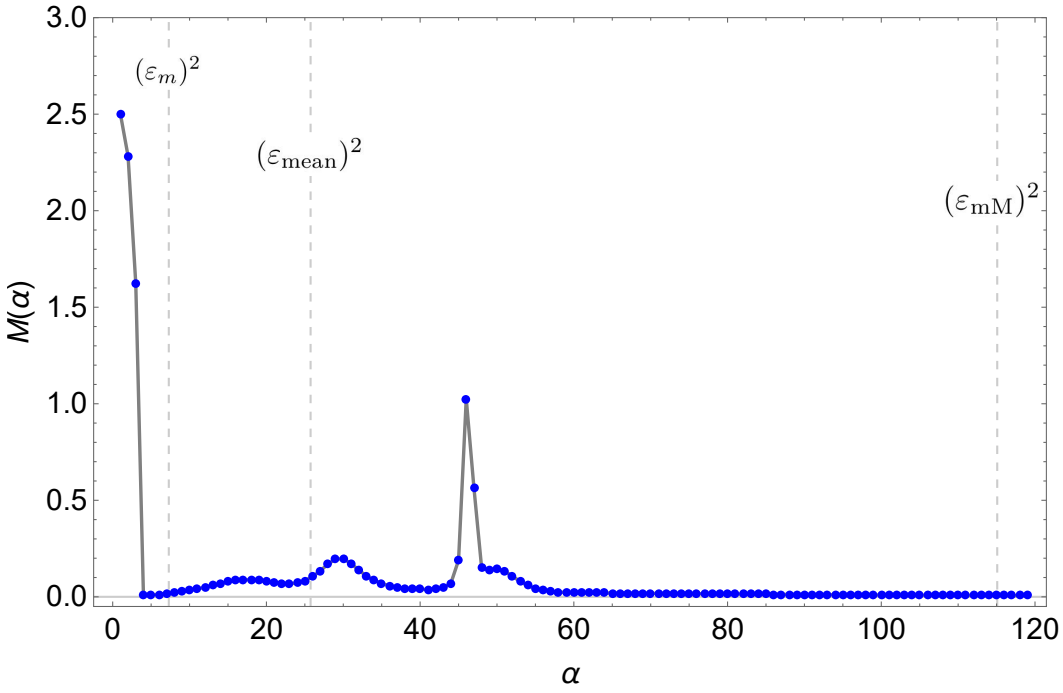
Figura 14 - Mapeamento 2D para diferentes valores do parâmetro $41 \leq \alpha \leq 55$.



Fonte: O autor, 2022.

Sobre $\alpha < 5$, a figura 15 exibe uma região desordenada nessa faixa. O mapeamento é impreciso, uma vez que o algoritmo perde precisão ao trabalhar com números muito pequenos. Por meio do link <https://youtu.be/D32zfJIeZ8w> é possível assistir o vídeo para essa faixa bem no início do vídeo. O exemplo evidencia que, entre os possíveis valores para o parâmetro de escala, existe um limite inferior onde, abaixo dele, o algoritmo não funciona ou não é confiável. Para o caso, este limite seria próximo de $\alpha = 5$.

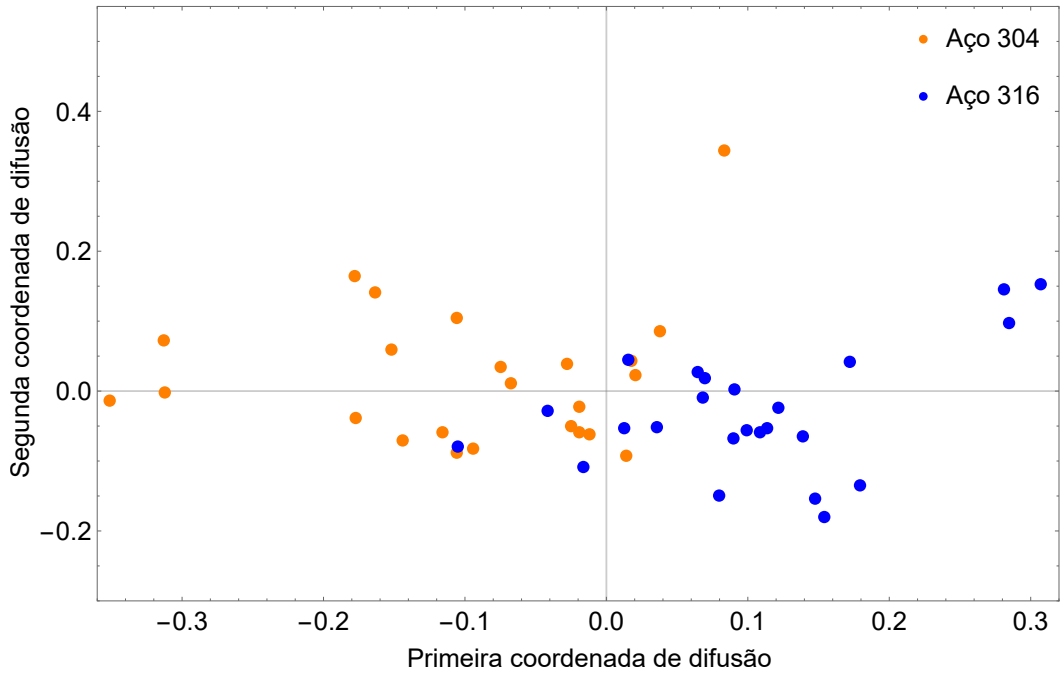
Figura 15 - $M(\alpha)$ para $1 \leq \alpha \leq 120$. Com valores menores que 5, aproximadamente, o parâmetro de escala torna o mapeamento impreciso devido à perda de precisão do algoritmo.



Fonte: O autor, 2022.

Diante da análise feita sobre o efeito do parâmetro α no comportamento do mapa de difusão em estudo, é razoável supor na escolha desse parâmetro com valor perto da região de estabilidade, próximo a $\alpha = 60$. No capítulo 5, com dados simulados, esta escolha será melhor justificada. Desta forma, as análises e resultados adiante para avaliar o parâmetro t utilizarão $\alpha = (\varepsilon_{\text{mM}})^2 = 115,085$ que, apesar de não estarem tão próximo dessa região para o valor do parâmetro α apresentado, levam a um mapeamento estável quando a separação eficiente desses perfis pela técnica tem melhor desempenho (Fig. 16).

Figura 16 - Mapeamento 2D obtido fazendo α igual ao quadrado do parâmetro de escala *min-max* do conjunto de dados



Fonte: O autor, 2022.

3.5 **Análise do parâmetro de escala temporal**

O próximo passo rumo à abordagem de classificação é avaliar o efeito do parâmetro temporal t no mapeamento dos dados por meio dos mapas de difusão. Como este parâmetro refere-se à potência da matriz de difusão, que por sua vez, carrega as probabilidades de difusão dos estados, espera-se conseguir uma análise geométrica multiescala do conjunto de dados. Cabe recordar que foi utilizado inicialmente $t = 2$ para a análise do parâmetro α.

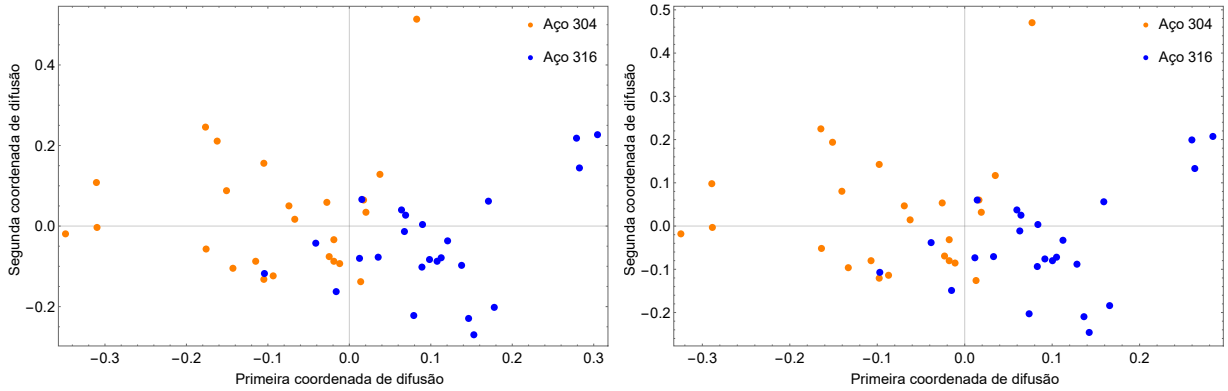
Fixando $\alpha = (\varepsilon_{\text{mM}})^2$, a figura 17 mostra o efeito do parâmetro t no mapeamento dos perfis usando novamente duas coordenadas principais (duas dimensões de classificação por perfil). Escolheu-se usar $t = 0,01$, $t = 0,1$, $t = 1$, $t = 5$, $t = 10$, $t = 50$ e $t = 100$, além de $t = 2$, usado anteriormente.

Ao aplicar os mapas de difusão e observar o efeito com a variação do parâmetro de escala temporal, verifica-se prontamente o impacto nas coordenadas de difusão. De fato, como mostra a equação 17, o parâmetro corresponde ao expoente dos autovalores que determinam as coordenadas de difusão dos dados no novo espaço e, como estes autovalores estão entre zero e um, suas potências decrescem com o aumento do parâmetro temporal, mas de forma diferente para cada autovalor. Logo, como espera-se que um processo de difusão deva ser, a concentração de informações acerca de um ponto de dado $\mathbf{x}_i$ específico

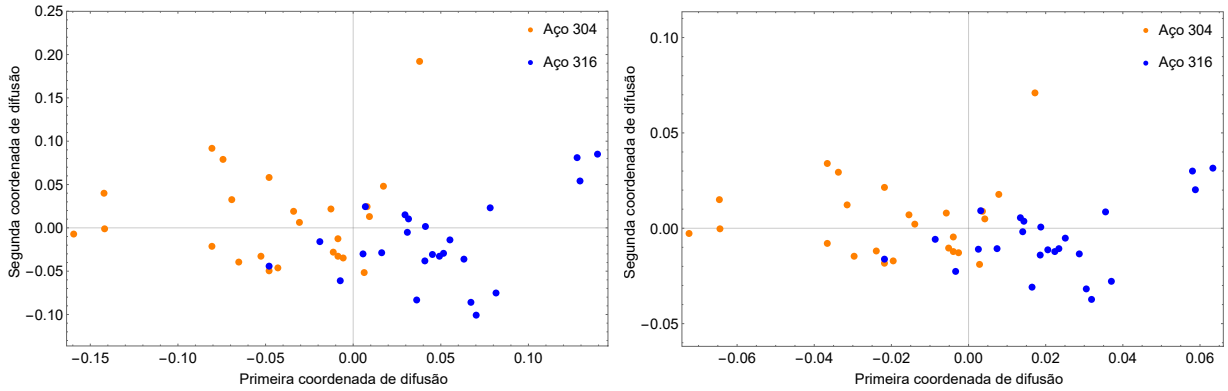
representado pelas entradas do vetor $\mathbf{z}_i$ que carrega as coordenadas desse ponto no novo espaço é diluída segundo a ordem decrescente de importância dos autovalores, fornecendo, em tese, ao decorrer da evolução do parâmetro, somente as coordenadas principais desse ponto de dados.

Em outras palavras, como o parâmetro t afeta diferentemente cada coordenada de difusão por meio da potência λ^t (levando todas elas de forma diferente a zero), o processo de difusão acontece de forma decrescente segundo a ordem de importância dos autovalores e, com isso, leva algumas destas componentes a zero (trazendo menos informações ao vetor ponto de dado mapeado—$\mathbf{z}_i$), contudo, teoricamente, dando as outras entradas restantes mais informações sobre o espaço original. As figuras 17a a 17h trazem o mapeamento 2D para diferentes valores do parâmetro temporal t com $\alpha = (\varepsilon_{\mathrm{mM}})^2$.

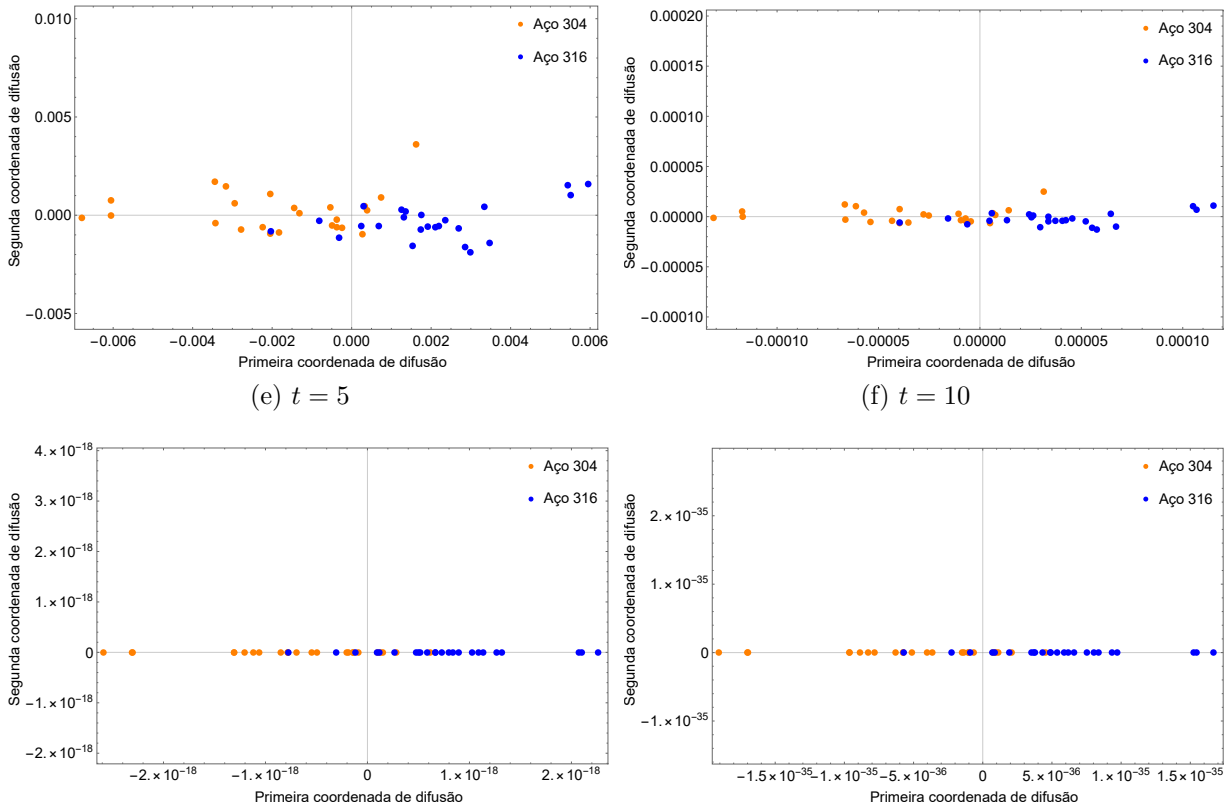
Figura 17 - Mapeamento 2D para diferentes valores do parâmetro t com $\alpha = (\varepsilon_{\mathrm{mM}})^2$.



(a) $t = 0,01$ (b) $t = 0,1$



(c) $t = 1$ (d) $t = 2$



(e) $t = 5$ (f) $t = 10$

(g) $t = 50$ (h) $t = 100$

Fonte: O autor, 2022

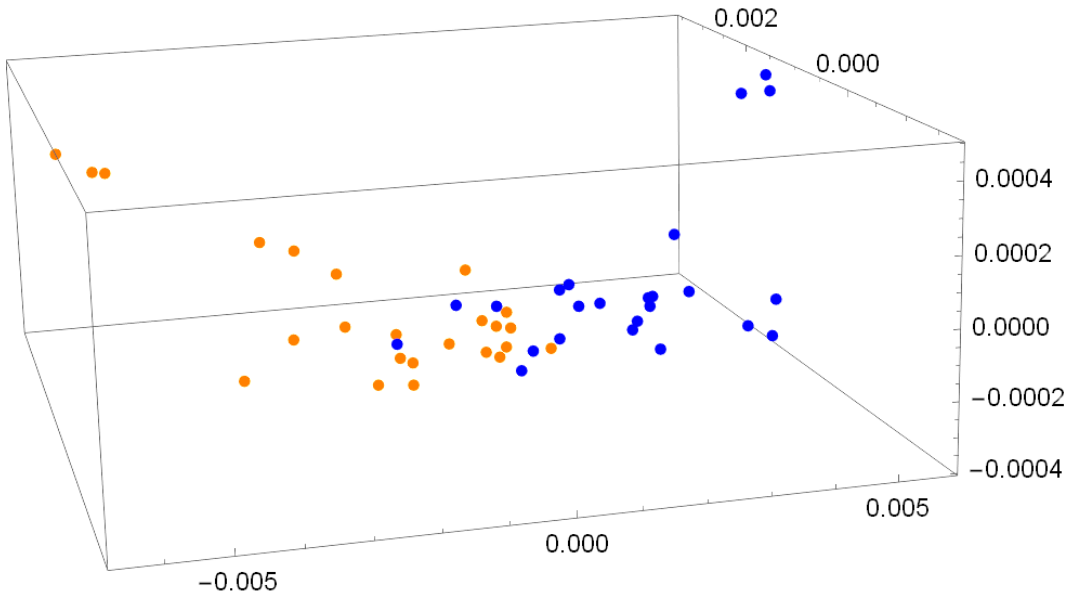
Ao verificar os diferentes mapeamentos presentes na figura 17, nota-se principalmente o efeito do parâmetro t na escala das coordenadas. As figuras nos extremos em 17a e 17h apresentam mapeamentos em diferentes escalas de magnitude, além de um achatamento vertical observado quando as janelas em cada figura trazem a escala de forma proporcional. Como observado anteriormente, o parâmetro afeta de forma diferente cada autovalor onde, cada um por sua vez, é responsável pelas diferentes coordenadas principais do mapeamento.

Em relação a disposição dos dados nos mapeamentos obtidos, concluiu-se que o impacto do parâmetro de escala é pouco significativo. De fato, pelo menos para valores não expressivos de t, os diferentes mapeamentos obtidos na figura 17 exibem basicamente a mesma distribuição geral dos dados no espaço de difusão sem distinção no arranjo dos pontos de dados mapeados, a não ser por distribuírem-se em regiões de escalas muito distintas. Uma vez que os métodos de separação se baseiam essencialmente nas distâncias relativas entre os pontos de dados, o efeito do parâmetro temporal para o objetivo de classificação dos dados é irrelevante e as diferentes taxas de acertos obtidas com o classificador *Bayes* são observadas somente com a alteração do parâmetro de escala α e não do parâmetro t. Por outro lado, não é recomendável assumir valores expressivos para o uso deste parâmetro (maior que 10, por exemplo, para o caso em estudo), principalmente para valores altos do parâmetro de escala α (quadrado do parâmetro de escala $min - max$

ou o diâmetro do conjunto de dados), uma vez que as coordenadas principais dos dados mapeados ficam representadas por números cada vez menores, prejudicando a qualidade das informações numéricas disponíveis em cada componente principal da redução que se objetiva.

Para as mesmas escolhas do parâmetro temporal t também foi feita a redução de dimensionalidade para três (mapeamento dos perfis usando três coordenadas principais). De igual forma, observou-se que o efeito desse parâmetro para o mapeamento 3D ocorre principalmente na escala. A figura 18 traz um dos mapeamentos usando $t = 5$. É possível perceber as duas regiões do espaço que contém cada grupo referente a cada um dos aços, contudo com alguma superposição dos pontos na região central da figura. Os pontos de dados mapeados com a técnica, mesmo neste espaço reduzido de poucas dimensões, parecem representar bem o sistema original e, assim como na figura 6 com os perfis das curvas de polarização, os dois grupos ficam bem definidos com a imersão.

Figura 18 - Mapeamento 3D obtido com a escolha do parâmetro de escala *min-max* do conjunto de dados e $t = 5$.



Fonte: O autor, 2022.

A existência de *outliers* nos perfis, juntamente com a forte não-linearidade das curvas de polarização dos diferentes aços envolvidos, torna árdua a tarefa do classificador na separação eficiente dos pontos de dados mapeados por meio dos mapas de difusão. É possível notar uma região de fracionamento onde, mesmo com uma efetiva técnica de separação, a tarefa de classificação correta dos aços parece ser trabalhosa. A seção seguinte traz os resultados obtidos com a técnica de classificação.

3.6 **Classificação**

Com o mapeamento realizado, o próximo passo foi implementar o classificador e avaliar os resultados. Para prever seu desempenho em novos dados, foi preciso avaliar sua taxa de erro em um conjunto de dados que não desempenhou nenhum papel no aprendizado do classificador. Este conjunto de dados independente é chamado de conjunto de teste. Presume-se que ambos os dados de treinamento e de teste sejam amostras representativas do problema subjacente (WITTEN, 2011).

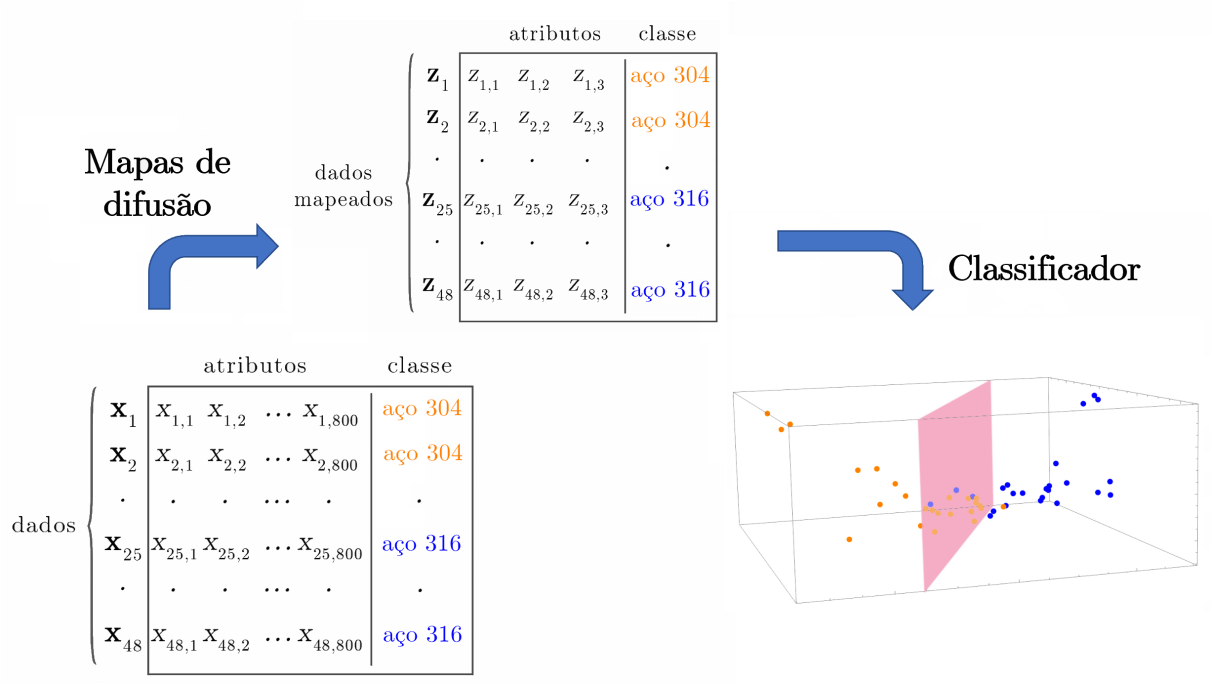
O conjunto de dados do perfil foi dividido aleatoriamente em 10 dobras, considerando que cada dobra contém perfis de ambos os aços, nas quais a classe é representada aproximadamente nas mesmas proporções que no conjunto de dados completo. Em cada execução deste esquema, o classificador é treinado usando todas menos uma dobra e, em seguida, é avaliado em como classifica as amostras da dobra separada. Esse processo se repete até que todas as dobras sejam usadas como conjunto de teste pelo menos uma vez e isso se justifica para tentar atenuar qualquer viés causado pela amostra específica escolhida para validação. As taxas de acerto nas diferentes iterações são calculadas para gerar uma taxa média de acerto geral. Este é o método de validação repetida da taxa de acerto.

Como é possível supor, uma única validação cruzada de 10 vezes pode não ser suficiente para obter uma estimativa de acerto confiável. De acordo com Witten (2011), diferentes experimentos de validação cruzada de 10 vezes com o mesmo método de aprendizagem e conjunto de dados frequentemente produzem resultados diferentes, por causa do efeito da aleatoriedade na escolha das próprias dobras. Dessa forma, a fim de obter uma estimativa de acerto mais precisa, o processo de validação cruzada é repetido até que todas as dobras sejam utilizadas, pelo menos, uma vez. Os resultados de classificação foram obtidos com o software *Weka* (AHER; LOBO, 2011) que faz essa rotina de forma automatizada. O esquema de redução de dimensionalidade e classificação dos dados das curvas de polarização é apresentado na figura 19.

A tabela 1 traz as taxas de acertos do classificador *Bayes* para diferentes escolhas do parâmetro α. Usa-se a abreviação Tc para taxa de acertos do classificador. Como a variação do parâmetro t não produziu diferenças entre as taxas de acerto em cada grupo, a tabela apresenta apenas a diferença quanto ao parâmetro de escala α utilizado. Indicado em negrito, a escolha de $\alpha = (\varepsilon_{\mathrm{d}})^2$ produziu os melhores resultados. O classificador acerta 39 dos perfis dentre os 48 e, com isso, produz uma taxa de acertos de $81,25\%$.

Da tabela 1, ainda é possível observar que a taxa de acerto obtida usando $\alpha = (\varepsilon_{\mathrm{d}})^2$ é muito próxima da taxa obtida usando $\alpha = (\varepsilon_{\mathrm{mM}})^2$, com a utilização do parâmetro de escala *min-max* do conjunto de dados. De fato, como o mapeamento obtido usando ambos os parâmetros exibe pouca diferença em relação à disposição dos dados mapeados (Fig 20), os resultados obtidos com o classificador são coerentes com os mapeamentos, uma vez que as figuras 20a e 20b exibem, praticamente, a mesma disposição dos dados no novo espaço.

Figura 19 - Esquema da abordagem de redução de dimensionalidade e classificação dos dados referentes às curvas de polarização.



Fonte: O autor, 2022.

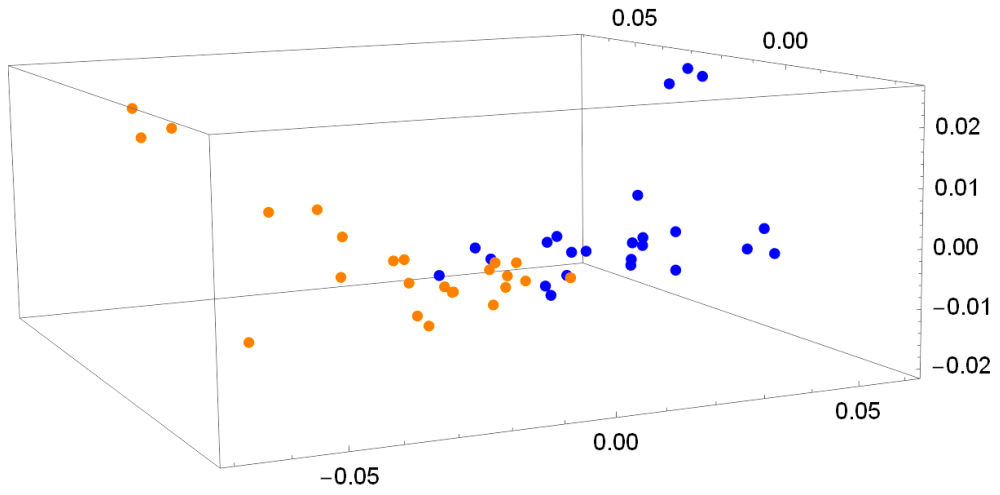
Tabela 1 - Taxas de classificação (Tc) para diferentes escolhas do parâmetro de escala α com o classificador *Bayes*. O número de perfis classificados corretamente (#PCC) para $\alpha = (\varepsilon_{\mathrm{d}})^2$ é 39.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{mean})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$\boldsymbol{(\varepsilon_{\mathrm{d}})^2}$
#PCC	26	30	38	**39**
Tc(%)	54, 16	62, 5	79, 17	**81,25**

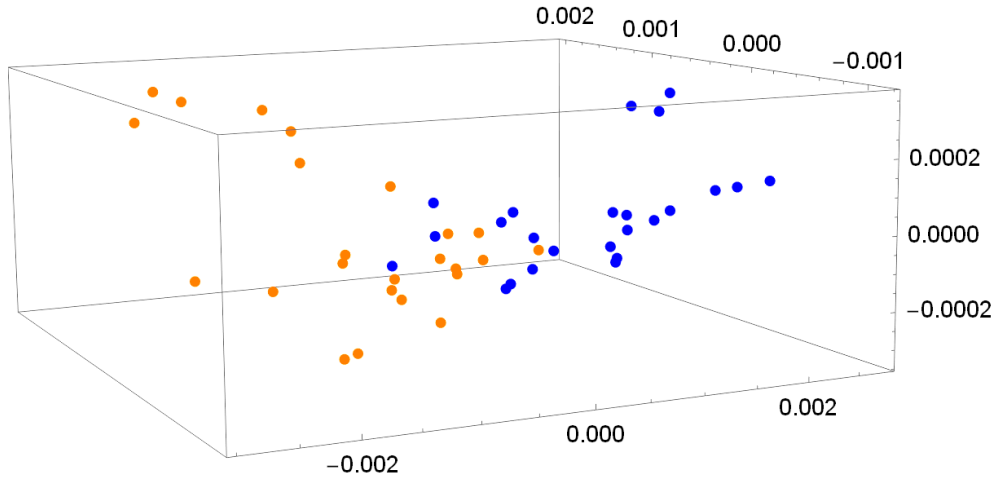
Fonte: O autor, 2022.

A seção seguinte traz a classificação realizada considerando separadamente para diferentes faixas de potencial eletroquímico aplicado.

Figura 20 - Mapeamentos 3D obtidos com $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $\alpha = (\varepsilon_{\mathrm{d}})^2$ para $t = 2$.



(a) $\varepsilon_{\mathrm{mM}}$



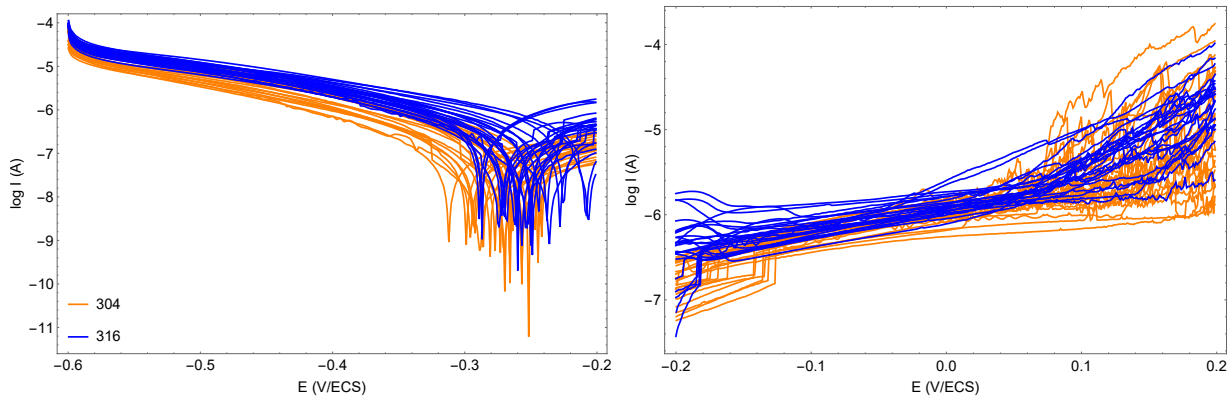
(b) ε_{d}

Fonte: O autor, 2022

3.6.1 Classificação separada de faixas de alto e baixo potencial

Assim como em Fabbri et al. (2014)—que utilizaram a abordagem multi-q em diferentes faixas de potencial de corrosão—também foram realizados experimentos de classificação para observar o comportamento da abordagem proposta em diferentes seções dos perfis que possuem propriedades distintas. Dessa forma, para cada perfil em cada faixa de estudo, foram utilizados vetores com 400 entradas (Fig. 21). Diferentes faixas de potencial estão relacionados à predominância de: (a) processos catódicos ($-0,6$ a $-0,2$ V $\times$ ECS), referentes a baixo potencial; e (b) anódicos ($-0,2$ a $0,2$ V $\times$ ECS), correspondente a alto potencial aplicado.

Figura 21 - Curvas de polarização dos aços 304 e 316 em duas diferentes faixas de potencial: (a)baixo potencial e (b)alto potencial



(a) (b)

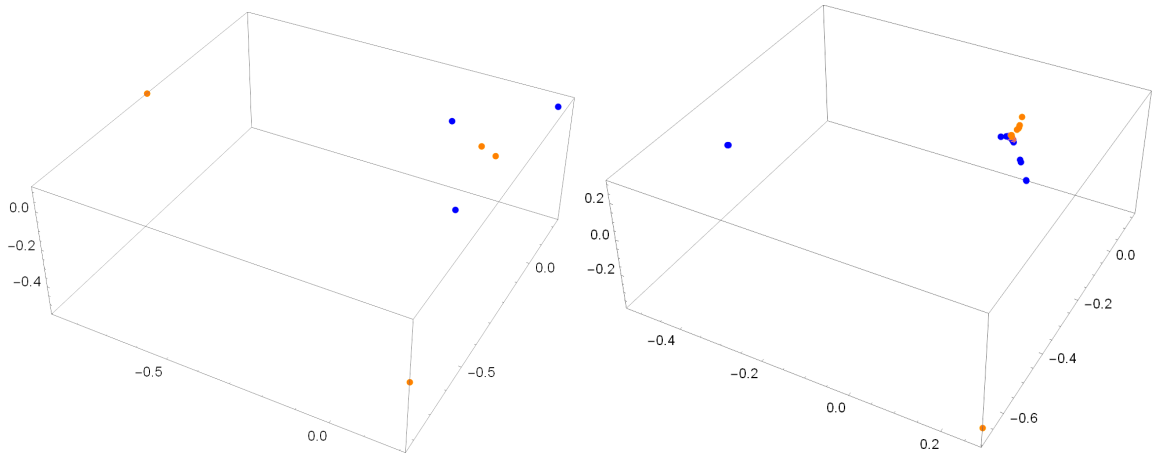
Fonte: O autor, 2022

3.6.1.1 Faixa de baixo potencial

Nesta subseção faz-se a análise dos perfis dos aços limitando agora a uma faixa específica de baixo potencial. Nesta faixa, os diferentes representantes de cada aço são mais facilmente identificados e, com isso, espera-se conseguir um mapeamento em que os *clusters* estejam melhor definidos.

Os perfis mapeados nesta faixa de potencial para diferentes α são mostrados na figura 22. Para a escolha do melhor valor deste parâmetro, fixou-se também $t = 2$. O resumo da taxa de classificação fica por conta da tabela 2. Como em Fabbri et al. (2014), ainda é difícil distinguir os perfis mesmo na faixa de baixo potencial. Neste sentido, ambos os métodos se comportaram de forma parecida.

Figura 22 - Mapeamento 3D com diferentes valores do parâmetro α referente à faixa de baixo potencial.



(a) $\alpha = (\varepsilon_{\mathrm{m}})^2$ (b) $\alpha = (\varepsilon_{\mathrm{mean}})^2$

(c) $\alpha = (\varepsilon_{\mathrm{mM}})^2$ (d) $\alpha = (\varepsilon_{\mathrm{d}})^2$

Fonte: O autor, 2022

Os mapeamentos expostos na figura 22 exibem os dados mapeados que se alteram significativamente com a mudança do parâmetro de escala α. Como é possível observar, os mapeamentos para $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $\alpha = (\varepsilon_{\mathrm{d}})^2$ agora têm diferenças expressivas, o que afeta a taxa de acerto obtida com o classificador.

Os resultados presentes na tabela 2 mostram que houve melhora significativa nas taxas de classificação em comparação aos resultados obtidos utilizando toda a faixa de potencial (Tab. 1). No melhor resultado encontrado, também em $\alpha = (\varepsilon_{\mathrm{d}})^2$, com $t = 2$, o classificador tem uma taxa de acertos de $87,5\%$ com 42 dos 48 perfis apresentados. Um ótimo resultado comparado com o obtido na faixa completa de potencial.

Tabela 2 - Taxas de classificação (Tc) para a faixa de baixo potencial para diferentes escolhas do parâmetro de escala α com o classificador *Bayes*.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{\mathrm{mean}})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$\boldsymbol{(\varepsilon_{\mathrm{d}})^2}$
#PCC	21	29	38	**42**
Tc(%)	$43,75$	$60,42$	$79,17$	**87,5**

Fonte: O autor, 2022.

3.6.1.2 Faixa de alto potencial

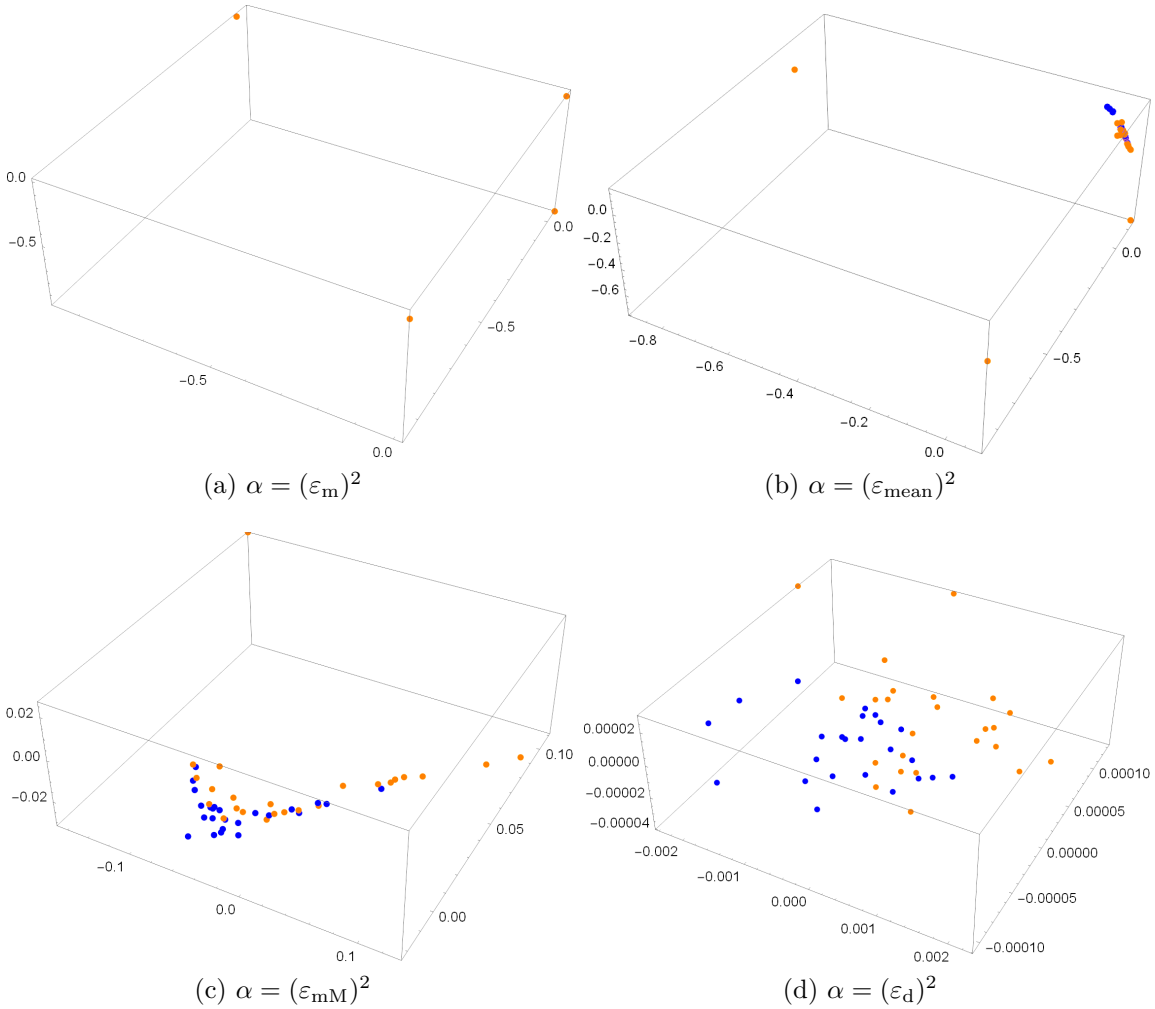
O comportamento eletroquímico sob polarização de dois aços inoxidáveis em solução aerada é semelhante, pelo menos na região catódica, e para potenciais logo acima do potencial de corrosão. Apesar de sua composição química diferente, principalmente em relação à porcentagem de molibdênio nos dois aços em estudo, uma sobreposição de curvas pode ser observada sob toda a faixa de potencial. Contudo, próximo ao potencial onde há maior suscetibilidade de corrosão, a diferença de resistência aparece com alguma dispersão e ocorre, ainda, um evento estocástico, que é a corrosão por pites. É esta a faixa de maior interesse no estudo de corrosão envolvendo os aços inoxidáveis.

Fazendo agora a análise dos perfis dos aços limitando à faixa específica de alto potencial, a figura 23 traz os perfis mapeados também para diferentes valores de α. Para a escolha do valor apropriado deste parâmetro, novamente fixou-se $t = 2$. O resumo da taxa de classificação é encontrada na tabela 3.

As taxas de classificação obtidas agora nesta faixa de potencial evidenciam uma dificuldade maior do classificador com a separação eficiente e a determinação correta dos rótulos dos aços. Na tabela 3, observando o melhor resultado encontrado usando $\alpha = (\varepsilon_{\mathrm{d}})^2$ e $t = 2$, o classificador tem uma taxa de acertos de $79,16\%$ com 38 dos 48 perfis apresentados. Apesar dessa taxa ser menor do que as taxas obtidas com a faixa completa

e de baixo potencial, é um ótimo resultado considerando a forte não-linearidade dos perfis nessa faixa. De fato, o percentual obtido nessa faixa foi, basicamente, o mesmo alcançado com a faixa completa de potencial.

Figura 23 - Mapeamento 3D com diferentes valores do parâmetro α na faixa de alto potencial.



(a) $\alpha = (\varepsilon_{\mathrm{m}})^2$ (b) $\alpha = (\varepsilon_{\mathrm{mean}})^2$

(c) $\alpha = (\varepsilon_{\mathrm{mM}})^2$ (d) $\alpha = (\varepsilon_{\mathrm{d}})^2$

Fonte: O autor, 2022

Tabela 3 - Taxas de classificação para a faixa de alto potencial para diferentes escolhas do parâmetro de escala α com o classificador *Bayes*.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{\mathrm{mean}})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$\boldsymbol{(\varepsilon_{\mathrm{d}})^2}$
#PCC	27	32	29	**38**
Tc(%)	56, 25	66, 67	60, 42	**79,17**

Fonte: O autor, 2022.

A seção seguinte traz a comparação dos resultados obtidos com os mapas de difusão e o classificador *Bayes* com outros métodos clássicos de redução de dimensionalidade, bem como com a abordagem multi-q, método utilizado em Fabbri et al. (2014) para obter o mapeamento dos dados no mesmo caso em estudo. Desta forma, é possível analisar estes diferentes métodos confrontando as vantagens e desvantagens de cada abordagem para os dados investigados.

3.7 **Comparação com outros métodos**

Nessa seção, a finalidade é comparar os resultados obtidos na seção anterior usando os mapas de difusão e o classificador *Bayes* com outros métodos na busca da separação eficiente dos perfis dos aços. Para cada método empregado, é feita a análise das taxas de classificação levando-se em conta as faixas de potencial de interesse e a complexidade do método. A melhor técnica, em tese, é aquela que consegue absorver melhor a estrutura presente no conjunto de dados e, com isso, conseguir uma redução significativa de dimensionalidade dos dados preservando suas características essenciais. Se isso ocorre, há um aumento na taxa de acertos obtida com o classificador.

3.7.1 Abordagem multi-q

O primeiro método apresentado para o estudo comparativo é o proposto em Fabbri et al. (2014) que usa a chamada abordagem multi-q baseada nas estatísticas de Tsallis para conseguir a redução e procede com o classificador *Bayes*. Nesse tratamento, o melhor resultado exposto traz uma taxa de classificação de 83% quando se leva em conta os perfis com toda a faixa de potencial (*full potential*), 90% quando se considera apenas a faixa de baixo potencial (*low potential*) e 80% com a faixa de alto potencial (*high potential*). Essa informação se encontra na linha em negrito na tabela 4, que é uma transcrição dos resultados apresentados nesse artigo.

Analisando as tabelas 1 a 3, em relação aos melhores resultados obtidos (também em negrito), conseguiu-se, respectivamente, 81%, 88% e 79% (taxa aproximada) contra 83%, 90% e 80% obtidos com a abordagem multi-q para essas faixas em seu melhor resultado (Multi-q, com $q = 0.1, 0.2, \ldots, 2.0$). Os mapas de difusão atingiram, basicamente, o mesmo resultado ao comparado com a taxa superior obtida com a abordagem multi-q com $q = 0.1, 0.2, \ldots, 2.0$, sendo superior a todos os outros listados na tabela 4.

Tabela 4 - Taxas de classificação (Tc) para diferentes métodos proposto por Fabbri et al. (2014) junto com seu melhor resultado em negrito.

Método	**Tc (%)**		
	Potencial completo	Baixo potencial	Alto potencial
Tsallis $q = 1$	83	83	65
Tsallis $q = 0.1$	73	65	60
Multi-q, $q = 0.1, 0.2, \ldots, 1.0$	73	69	69
Multi-q, $q = 0.1, 0.2, \ldots, 2.0$	**83**	**90**	**80**
Bayes em todos os 800 pontos	81	75	73

Fonte: O autor, 2022.

A diferença entre os resultados apresentados com o melhor resultado dos dois métodos (mapas de difusão e multi-q, $q = 0.1, 0.2, \ldots, 2.0$) foi de apenas 1 perfil. Na faixa completa de potencial, os mapas de difusão obtiveram uma taxa de acerto de 39 contra 40 dos perfis tratados com multi-q, e 42 contra 43 no baixo potencial. Na faixa de alto potencial, aonde no geral se tem o maior interesse, ambos os métodos obtiveram o mesmo resultado. Isso evidencia a capacidade dos mapas de difusão na eficiente redução de dimensionalidade não-linear, assim como a abordagem multi-q, configurando robustos métodos para o estudo e análise de sinais eletroquímicos.

Também é possível concluir que ambos os métodos são coerentes em relação à taxa de acertos de classificação para as faixas de potencial em estudo em relação à dificuldade de classificação. Os resultados obtidos com os mapas de difusão mostram que, assim como em Fabbri et al. (2014), os perfis exibem significativamente mais desafios para classificação nas faixas de alto potencial, sugerindo, então, que a região de estudo contemple ambas as faixas para melhor desempenho (faixa completa ou total do potencial).

3.7.2 PCA - Análise de Componentes Principais

O método exposto brevemente na seção 2 também foi implementado e aplicado aos dados nesse trabalho. As figuras 24 a 26 trazem o mapeamento para três coordenadas principais com as faixas completa, baixo e alto potencial, respectivamente.

Apesar de fácil implementação, os mapeamentos mostram exatamente o que o capítulo 1 trouxe: a PCA não conseguiu absorver as características não-lineares dos perfis. Os mapeamentos dos dados, principalmente na faixa alto potencial, não sugerem uma separação e, com isto, o classificador pouco consegue êxito. Apesar de atingir um ótimo

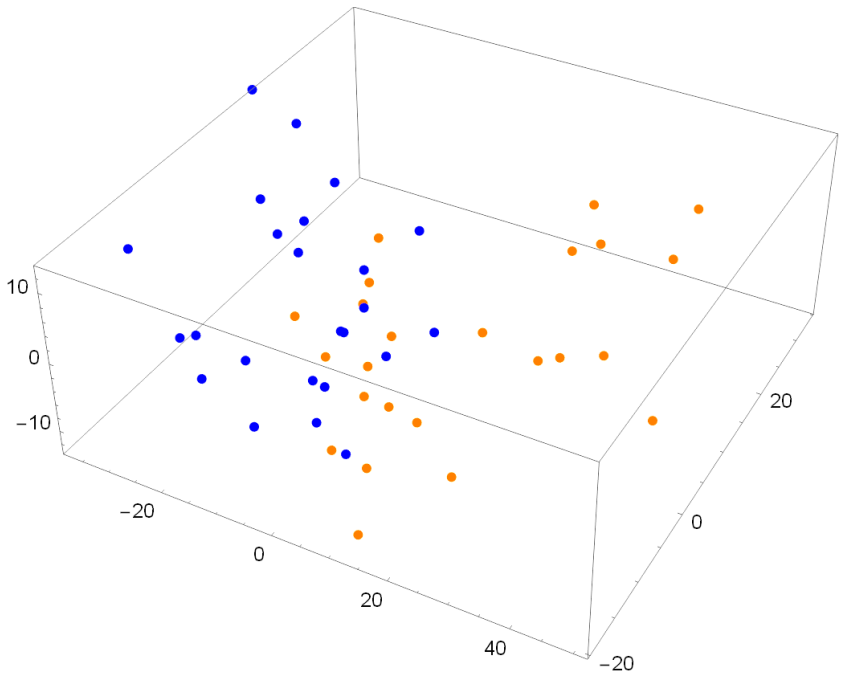
resultado na faixa de baixo potencial, talvez devido a aparência mais linear dos perfis nessa faixa, os resultados da técnica contidos na tabela 5 evidenciam, principalmente para faixa de alto potencial, que a técnica não é adequada para o fim proposto.

Tabela 5 - Taxas de classificação (Tc) para as distintas faixas de potencial utilizando a abordagem clássica de Análise de Componentes Principais (PCA).

	Faixa de potencial		
	Completo	Baixo	Alto
#PCC	36	43	28
Tc(%)	75	90	58

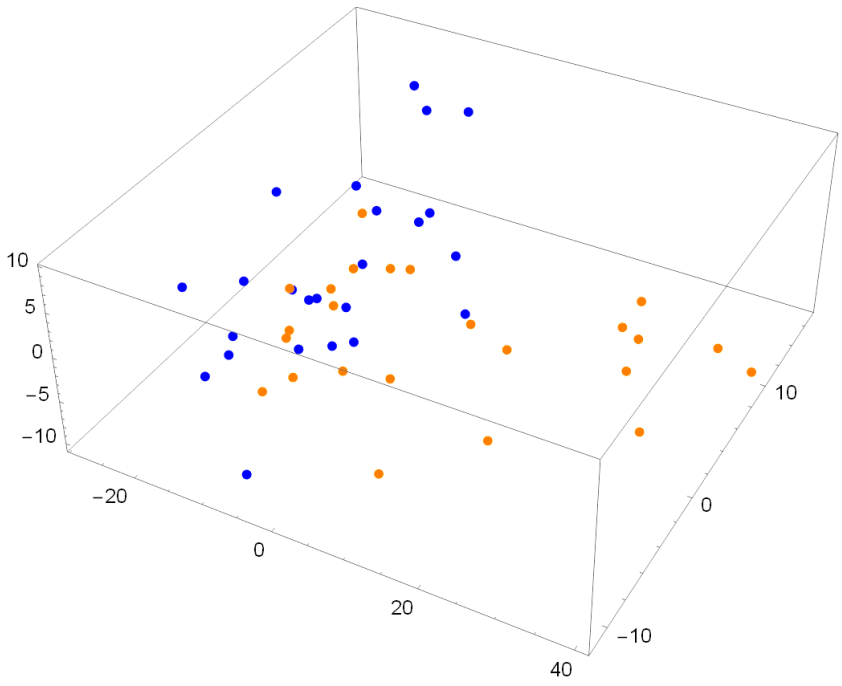
Fonte: O autor, 2022.

Figura 24 - Análise clássica de componentes principais 3D nos dados de perfil referentes à faixa completa de potencial.



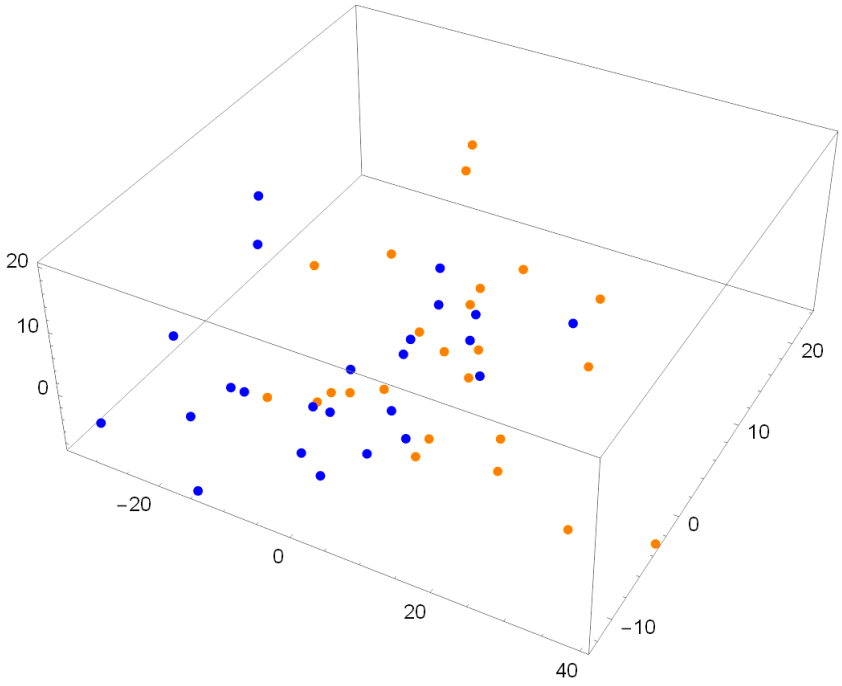
Fonte: O autor, 2022.

Figura 25 - Análise clássica de componentes principais 3D nos dados de perfil referentes à faixa de baixo potencial.



Fonte: O autor, 2022.

Figura 26 - Análise clássica de componentes principais 3D nos dados de perfil referentes à faixa de alto potencial.



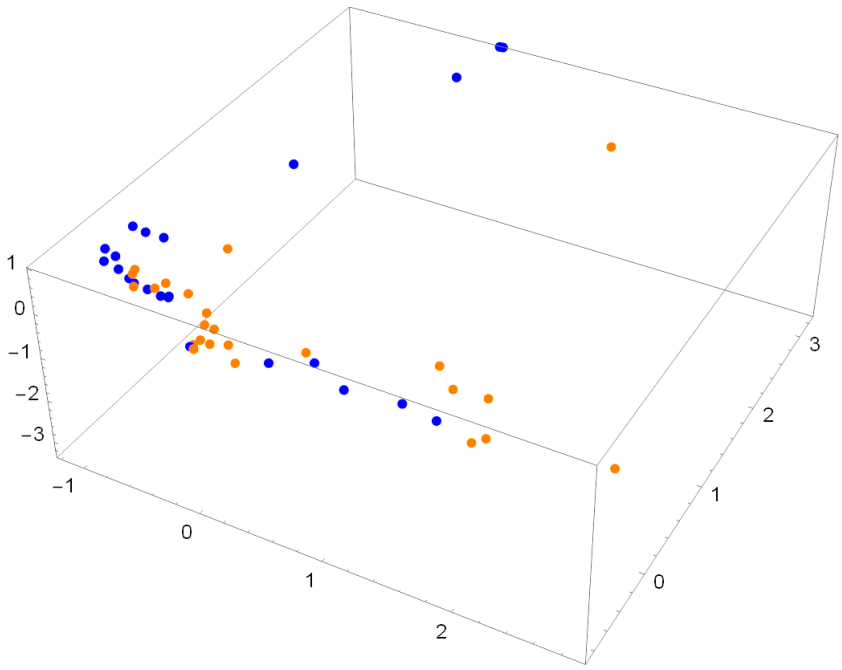
Fonte: O autor, 2022.

3.7.3 Imersão Localmente Linear

Cabe a essa subseção mostrar também os resultados obtidos pela técnica de imersão localmente linear (Locally Linear Embedding - LLE). Como abordou o capítulo 1, a técnica tem possibilidade de considerar a não-linearidade da estrutura e, com isso, supostamente alcançar melhores resultados do que o método PCA, por exemplo, mostrado na seção anterior. As figuras 27 a 29 trazem o mapeamento para três coordenadas principais com as faixas de completo, baixo e alto potencial, respectivamente, agora para tal abordagem.

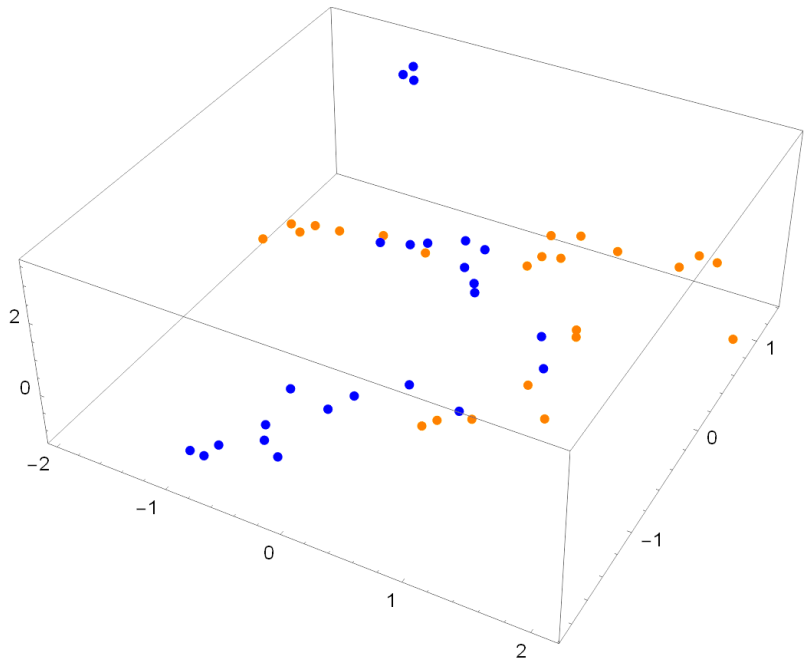
Apesar da técnica dessa vez ter a capacidade para lidar com dados inseridos em um espaço não-linear, o mapeamento obtido com o método mostra que, a sobreposição dos perfis e a presença de *outliers* em cada aço, dificulta que esse seja feito de forma eficiente, principalmente na faixa completa e de alto potencial. No mapeamento de completo potencial (Fig. 27), por exemplo, os dados mapeados estão quase todos aglomerados sobre uma região, onde é possível notar sobreposição nessa faixa, dificultando posteriormente a tarefa do classificador.

Figura 27 - Mapeamento 3D obtido com o método LLE nos dados de perfil referentes faixa completa de potencial.



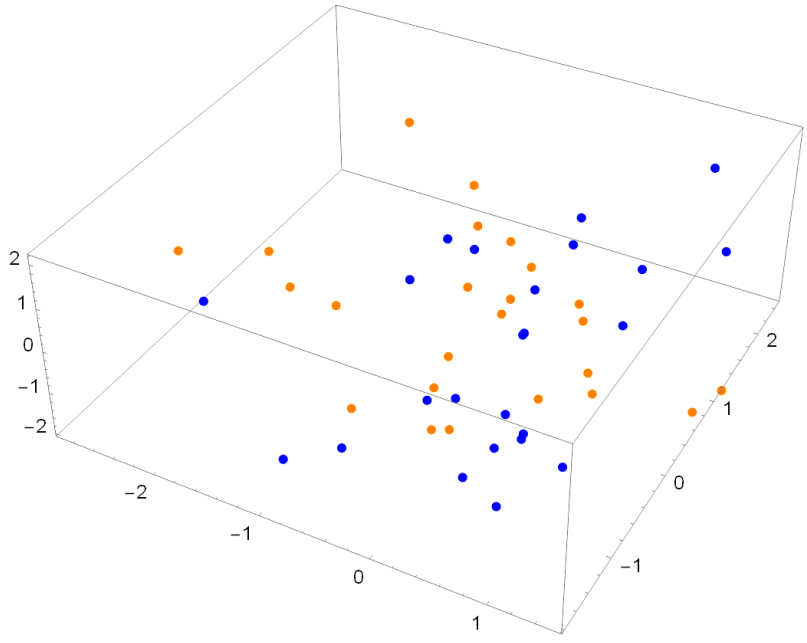
Fonte: O autor, 2022.

Figura 28 - Mapeamento 3D obtido com o LLE nos dados de perfil referentes à faixa de baixo potencial.



Fonte: O autor, 2022.

Figura 29 - Mapeamento 3D obtido com o LLE nos dados de perfil referentes à faixa de alto potencial.



Fonte: O autor, 2022.

Os resultados da classificação obtidos nos dados mapeados com a técnica estão contidos na tabela 6. É possível observar que as taxas de classificação obtidas são condizentes com o que a técnica propõe e os resultados obtidos são tão bons quanto várias outras exibidas na tabela 4. Por outro lado, os resultados com o LLE não foram melhor do que o *Bayes* em todos os 800 pontos para a faixa de baixo potencial, onde não foi realizada nenhuma redução de dimensionalidade.

Tabela 6 - Taxas de classificação (Tc) para as distintas faixas de potencial utilizando o método de Imersão Localmente Linear (LLE).

	Faixa de potencial		
	Completo	Baixo	Alto
#PCC	36	36	35
Tc(%)	75	75	73

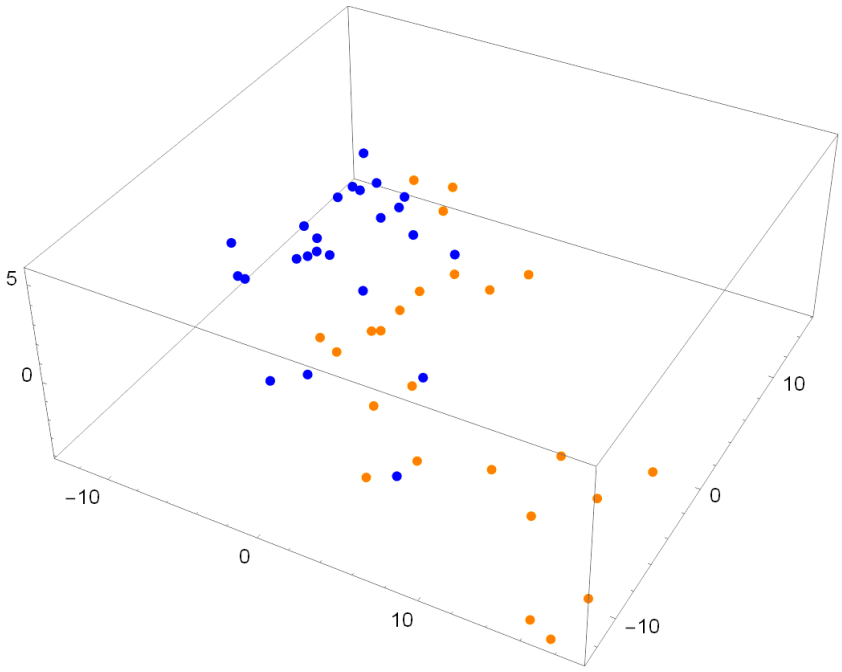
Fonte: O autor, 2022.

3.7.4 Mapas de características isométricas

Como último método para fim de comparação, nessa subseção são trazidos também os resultados obtidos pelo técnica de *isomap*. Partindo da ideia que ela generaliza a técnica de dimensionamento multidimensional (MDS) para alcançar a redução, os resultados com o MDS não foram realizados. As figuras 30 a 32 trazem o mapeamento para três coordenadas principais com as faixas de completo, baixo e alto potencial, respectivamente.

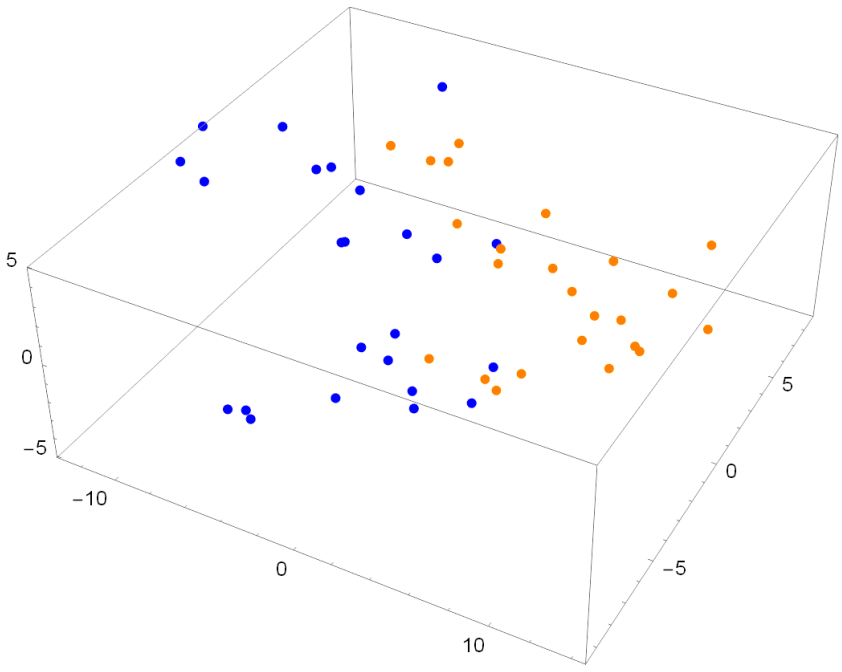
Os resultados da técnica contidos na tabela 7 mostram também que ela faz jus à sua reputação. As taxas de classificação obtidas com o método, em todas as faixas contempladas, são ainda melhores do que o LLE e isso também a coloca em destaque diante dos métodos apresentados. Tem um boa resposta para a faixa de alto potencial, o que a coloca dentre as mais indicadas nessa faixa. Em contrapartida, não tem um desempenho em tal grau na faixa de baixo potencial colocando-a em uma posição de aproveitamento menor do que o multi-q, $q = 0.1, 0.2, \ldots, 2.0$ e os mapas de difusão.

Figura 30 - Mapeamento 3D obtido com o *isomap* nos dados de perfil referentes à faixa completa de potencial.



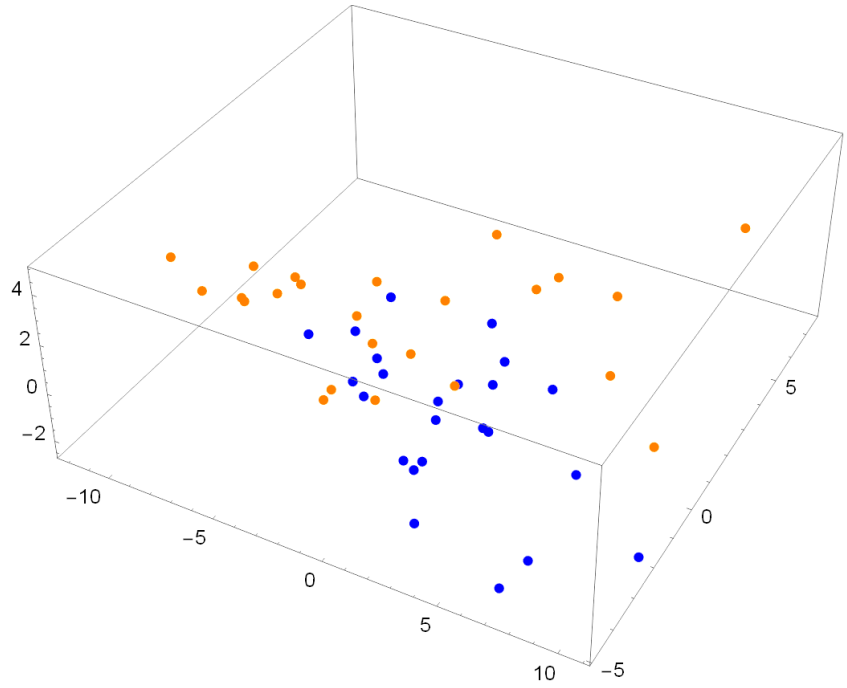
Fonte: O autor, 2022.

Figura 31 - Mapeamento 3D obtido com o *isomap* nos dados de perfil referentes à faixa de baixo potencial.



Fonte: O autor, 2022.

Figura 32 - Mapeamento 3D obtido com o *isomap* nos dados de perfil referentes à faixa de alto potencial.



Fonte: O autor, 2022.

Tabela 7 - Taxas de classificação (Tc) para as distintas faixas de potencial utilizando o método *isomap*.

	Faixa de potencial		
	Completo	Baixo	Alto
#PCC	39	37	37
Tc(%)	81	77	77

Fonte: O autor, 2022.

3.7.5 Resumo

Para sintetizar todos os resultados obtidos diante de cada faixa de potencial analisada com os métodos apresentados, esta subseção final traz uma tabela resumo das taxas de classificação obtidas. Dessa forma, o objetivo da tabela 8 é apresentar a comparação final do desempenho de cada método discutido exibindo a taxa de classificação média dos aços.

Tabela 8 - Resumo das taxas de classificação obtidas para os diversos métodos apresentados.

Método	**Taxa de classificação (%)**		
	Potencial completo	Baixo potencial	Alto potencial
Tsallis $q = 1$	83	83	65
Tsallis $q = 0,1$	73	65	60
Multi-q, $q = 0,1; 0,2; \ldots; 1,0$	73	69	69
Multi-q, $q = 0,1; 0.2; \ldots; 2,0$	83	90	80
Bayes em todos os 800 pontos	81	75	73
Mapas de difusão	81	88	79
PCA	75	90	58
LLE	75	75	73
isomap	81	77	77

Fonte: O autor, 2022.

Com base no que foi apresentado, vê-se que os mapas de difusão se configuram como uma técnica robusta e promissora também para a análise de perfis de corrosão derivados de sinais eletroquímicos. Apesar de não superar o multi-q no melhor do seu desempenho (apenas um perfil classificado incorretamente de diferença), a técnica mostrou-se superior aos métodos Tsallis com $q = 0,1$ e multi-q com $q = 0,1; 0,2; \ldots; 1,0$ em todas as faixas, conseguindo também melhores resultados comparado ao Tsallis com $q = 1$ na faixa de alto e baixo potencial e ao *Bayes* sem redução em todas as faixas. Ainda, conseguiu resultados superiores em comparação a todos os métodos clássicos de redução de dimensionalidade apresentados como a PCA, o LLE e o *isomap*.

É possível também concluir que a faixa de alto potencial é mesmo um desafio para os métodos testados. Nesta faixa onde o processo de corrosão se acentua, a diferença de resistência aparece junto com alguma dispersão estocástica, o que resulta em um comportamento do perfil quase que aleatório. No capítulo seguinte, é mostrado que uma solução para melhorar os resultados obtidos de uma classificação eficiente pode ser encontrar e excluir *outliers* no conjunto de dados. Esse procedimento também é realizado com êxito com os mapas de difusão.

4 MAPAS DE DIFUSÃO NA BUSCA DE *OUTLIERS* EM SINAIS ELETROQUÍMICOS

Os procedimentos sistemáticos realizados para obter dados experimentais acerca do caso em estudo, como o descrito no capítulo anterior, fazem parte da rotina de pesquisadores na busca do entendimento sobre o funcionamento mais profundo dos fenômenos envolvidos e as relações entre as variáveis presentes nestes processos.

Ao dispor desses dados experimentais, naturalmente é de se imaginar que eles sejam compostos de uma parcela de ruído, decorrente da experimentação, e que, eventualmente, algumas observações podem não traduzir de forma adequada o processo investigado. No caso de estudo abordado nesse trabalho, por exemplo, o comportamento dos aços em estudo em relação à corrosão localizada pode não ser bem representado por algumas das medições efetuadas. A esses perfis, doravante, refere-se como *outliers* do conjunto de dados.

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 à medida que serve de modelo para o emprego de ferramentas de monitoramento estatístico nas indústrias de manufatura e serviços. Segundo Wang et al. (2015), o monitoramento do processo é uma questão importante para garantir a segurança e operações eficientes nas indústrias de processo modernas.

O objetivo do capítulo é duplo. Primeiramente, busca-se por meio da aplicação da técnica de mapas de difusão encontrar *outliers*, fato que, aparentemente, possibilitaria um ganho na classificação em relação a todo o conjunto de dados. Se for possível identificar inicialmente os perfis discrepantes para o conjunto de dados experimentais do capítulo anterior, pode-se suprimi-los e reavaliar a classificação para o novo conjunto contendo apenas dados válidos. A esperança é que haja uma depuração dos dados e que isso implique em melhores resultados com o classificador. Adicionalmente, é apresentado também um procedimento de pré-processamento dos dados disponíveis por meio da lei de Benford.

Outro objetivo é a aplicação da técnica em perfis de sinais de impedância eletroquímica. Diante de um novo conjunto de dados que agora exibe os perfis de impedância de outro aço conhecido, o objetivo é também conseguir a depuração destes dados e discutir a eficiência da técnica em outro exemplo envolvendo também dados experimentais. Os mapas de difusão na busca de *outliers* foram utilizados com sucesso, por exemplo, em perfis provenientes da produção de madeira processada em relação à massa específica ao longo da espessura (MOURA NETO; SOUZA; MAGALHÃES, 2019).

4.1 **Lei dos dígitos significativos como um pré-processamento**

Antes de exibir como os mapas de difusão podem também ser úteis na busca de *outliers* em um conjunto de dados, o propósito desta seção é mostrar o curioso ajuste da lei dos dígitos significativos aos dados de estudo e, de outra forma, permitir fornecer um tipo de selo de autenticidade aos dados experimentais utilizados neste trabalho. Isso posto, a lei dos dígitos significativos, aqui brevemente abordada, pode servir como um pré-processamento para a análise dos dados, verificando a consonância da distribuição dos dados obtidos experimentalmente com a distribuição teórica logarítmica. Após uma breve apresentação da referida lei, discute-se a sua aplicação no contexto do estudo de sinais eletroquímicos.

4.1.1 Lei de Benford

A lei dos dígitos significativos, ou lei dos números anômalos, como também é conhecida, recebeu o nome de lei de Benford em homenagem ao engenheiro elétrico e físico americano Frank Benford ($1883 - 1948$). Em 1938, estimulado pelas pistas deixadas pelo astrônomo e matemático Simon Newcomb ($1835 - 1909$), Benford publicou o artigo intitulado *The law of anomalous numbers* (BENFORD, 1938), onde mostrava que, diferentemente do que o senso comum possa indicar, a frequência do dígito inicial de uma coleção de números aleatórios segue uma distribuição nada uniforme com mais dígitos iniciais começados com 1 do que com 9.

Em 1881, Newcomb já tinha reportado sobre suas observações e, além disto, proposto uma distribuição de frequências para os primeiro e segundo dígitos. De acordo com este relato, as probabilidades necessárias de ocorrência no caso dos dois primeiros dígitos estão representadas na tabela 9, para o sistema numérico na base 10. É interessante observar que, a partir do segundo dígito, a distribuição se aproxima da distribuição uniforme, com probabilidade aproximada de $0,1$ para cada dígito.

Tabela 9 - Distribuição teórica da lei de Benford para os dois primeiros dígitos.

Dígito	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1º	-	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
2º	0,112	0,114	0,109	0,104	0,1	0.097	0,093	0,09	0,088	0,085

Fonte: O autor, 2022.

Segundo Benford (1938), foram coletados dezenas de milhares de números de diferentes fontes e domínios e mostrou-se que todos seguiam a mesma distribuição. Dentre

os dados analisados, estavam tamanhos de populações de diferentes localidades, áreas de superfície de rios, constantes físicas, massas moleculares, endereços diversos, taxas de natalidade e mortalidade e até estatísticas de jogos de beisebol. Ainda que há quem acredite e possa oferecer evidências convincentes de que Benford manipulou erros de arredondamento para obter um melhor ajuste à lei logarítmica, segundo Hill (1995), no entanto, mesmo os dados não manipulados têm um bom ajuste.

No decorrer dos anos que se seguiram, este comportamento foi verificado para muitos outros dados. A partir da década de 1990, principalmente por conta do avanço computacional, a lei de Benford destacou-se no campo da auditoria com foco nos registros contábeis (COSTA, 2010). Aplicado aos balanços financeiros da empresa *Enron Corporation* nos anos de 2001 e 2002, o método apontou um aumento anormal da receita no período da fraude, indicando que a empresa teria inflado artificialmente seus lucros. O caso ganhou destaque na época, dada a enorme relevância econômica desta empresa de energia nos Estados Unidos, ganhando manchete nos jornais. Por este e outros exemplos, o método provou ser uma valiosa ferramenta para detecção de manipulação, fraude e alterações nas demonstrações financeiras entre outros contextos (NIGRINI; MITTERMAIER, 1997).

4.1.2 Formulação matemática

A lei dos dígitos significativos decorre de observação empírica. Os dados numéricos reais apresentam uma distribuição dos algarismos significativos iniciais que não é uniforme, diferentemente de como talvez aponte o senso comum, mas obedecem a uma distribuição logarítmica específica.

Seja $P(D_i = s)$ a probabilidade do i-ésimo dígito significativo ser s. Se $i = 1$, tem-se a probabilidade associada ao dígito inicial (primeiro dígito significativo). De acordo com Newcomb (1881),

$$P(D_1 = s) = \log\left(1 + \frac{1}{s}\right) \tag{30}$$

e

$$P(D_2 = s) = \sum_{i=1}^{9} \log(1 + (10i + s)^{-1}) \tag{31}$$

A tabela 9 exibe os valores de $P(D_1 = s)$ e $P(D_2 = s)$ dados pelas equações 30 e 31.

Em Hill (1995), além das leis para o primeiro e segundo dígito, também é apresentada a lei geral dos dígitos significativos ou a distribuição de probabilidade conjunta dos i

primeiros dígitos:

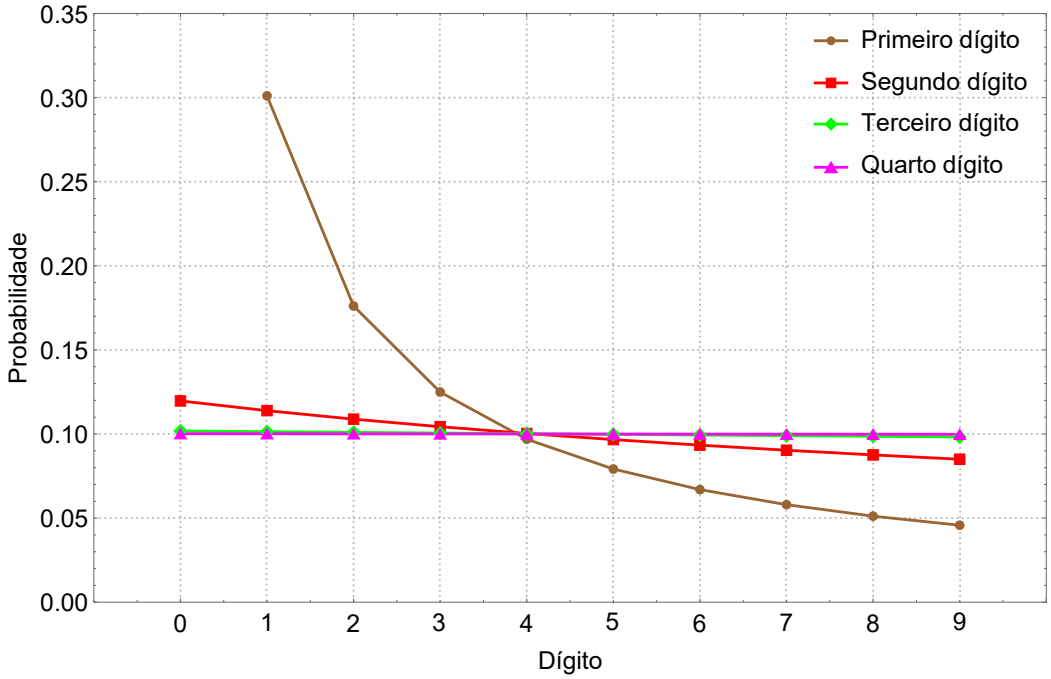
$$P(D_1 = s_1, D_2 = s_2, \ldots, D_i = s_i) = \log\left[1 + \left(\sum_{j=1}^{i} s_j \cdot 10^{i-j}\right)^{-1}\right]. \quad (32)$$

Observe que a equação 30 é um caso particular da equação 32 levando em conta que s_i pode assumir todos os valores de 1 a 9, e que, a partir da equação 32, é possível obter-se a equação 31, considerando a marginal,

$$P(D_2 = s) = \sum_{j=1}^{9} P(D_1 = j, D_2 = s). \quad (33)$$

Outro fato interessante evidenciado na equação 32 é que, ao contrário do que se poderia esperar, os dígitos significativos de um número qualquer são dependentes (HILL, 1995). Isso significa que, assim como os primeiros seguem uma distribuição bem definida, o segundo também depende do primeiro, o terceiro do segundo e do primeiro, e assim por diante. É interessante observar que, em muitos dados experimentais, este comportamento se verifica. À medida que a distância entre os dígitos de interesse aumenta, a dependência entre os dígitos significativos diminui rapidamente, tendendo para a distribuição uniforme. Em seu artigo, apesar de não apresentar tais probabilidades, Newcomb (1881) já havia percebido que as probabilidades relacionadas ao terceiro e quarto dígito se aproximavam da distribuição uniforme. A figura 33 mostra a frequência dos quatro primeiros dígitos segundo a equação 32. A partir do segundo dígito, a distribuição é aproximadamente uniforme, com probabilidade de $0,10$ para cada dígito. Isto evidencia, de fato, que a distribuição relacionada ao primeiro dígito é a mais interessante para fins de monitoramento do comportamento natural ou daquele sob manipulação.

Figura 33 - Distribuição teórica de frequências dos quatro primeiros dígitos.



Fonte: O autor, 2022.

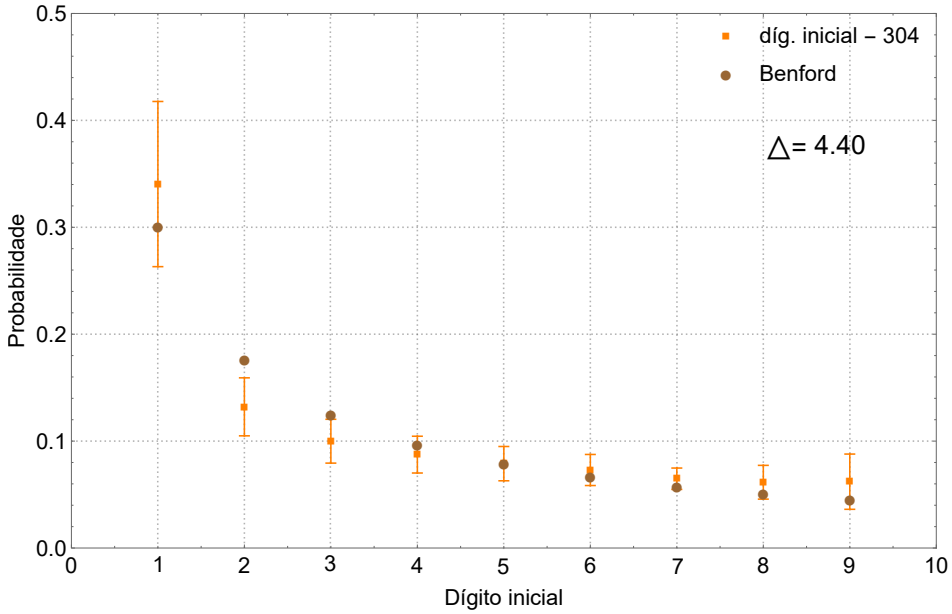
4.1.3 Curvas de polarização e a lei de Benford

Nesta subseção é apresentada a distribuição dos dígitos referentes aos dados eletroquímicos. O procedimento é bem simples, e consistiu em obter o dígito significativo em cada um dos números que compõem o banco de dados que representa cada uma das grandezas obtidas por meio da experimentação.

Para os dados referentes às curvas de polarização, a lei de Benford foi aplicada a cada um dos 48 perfis. A fim de analisar o efeito do tipo de aço e de modo geral, obteve-se a média em cada grupo e o desvio-padrão. Cabe acrescentar que os dados utilizados se referem à medida da corrente observada para cada potencial utilizado antes da aplicação do logaritmo. As figuras 34 e 35 apresentam os resultados para o primeiro dígito significativo junto com a distribuição de Benford dos dados da curva de polarização para cada um dos aços inoxidáveis.

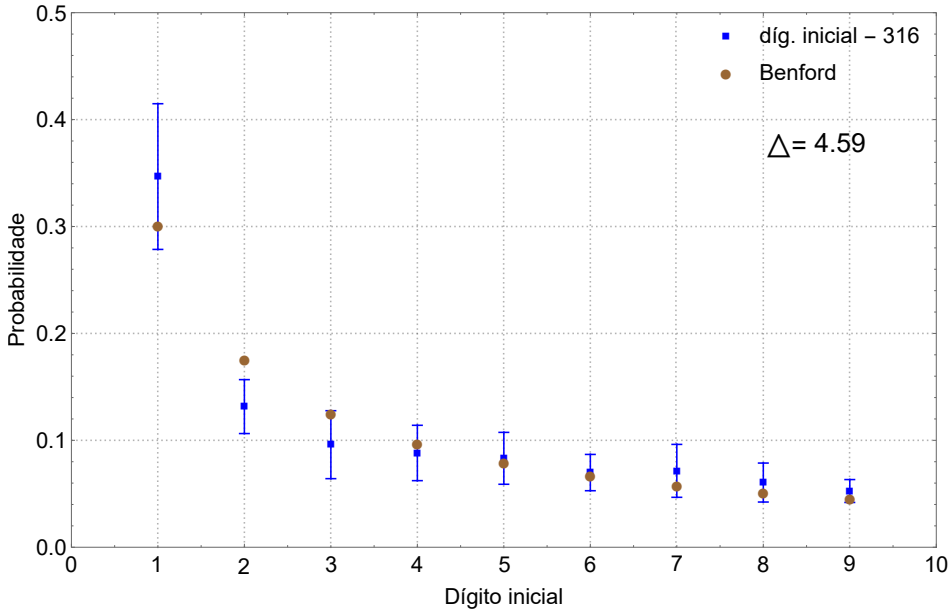
Uma rápida inspeção visual inicialmente permite suspeitar a conformidade dos dados com a lei. Curiosamente, a distribuição logarítmica de Benford parece modelar também aqui os dados obtidos com as parcelas de corrente e as probabilidades associadas a cada dígito ficam dentro da margem de erro para, praticamente, todos os dígitos. Somente no dígito 2 e com o 3, para o aço 304, a porcentagem relativa à distribuição teórica fica fora da margem de erro dada pelo desvio padrão. Ainda assim, pode-se observar um bom ajuste entre as duas distribuições.

Figura 34 - Distribuição média dos primeiros dígitos dos 24 perfis do aço 304 com seu desvio-padrão (laranja) e a distribuição segundo a lei de Benford (marrom) da corrente em valores absolutos.



Fonte: O autor, 2022.

Figura 35 - Distribuição média dos primeiros dígitos dos 24 perfis do aço 316 com seu desvio-padrão (azul) e a distribuição segundo Benford (marrom) da corrente em valores absolutos.



Fonte: O autor, 2022.

A fim de apresentar uma medida de quão próximo as distribuições obtidas estão das distribuições de Benford (teste de aderência), é exibido nos gráficos de comparação das distribuições o parâmetro Δ.

$$\Delta = 100 \cdot \max_{s=1}^{9} \left| P(D_1 = s) - \log\left(1 + \frac{1}{s}\right) \right| \tag{34}$$

Assim como em (HILL, 1995), trata-se simplesmente da diferença máxima, em percentual, entre as probabilidades dos primeiros dígitos significativos de uma distribuição dada e as probabilidades dadas pela equação 32. Desse modo, por exemplo, $\Delta = 0$ indica a exata conformidade de uma distribuição analisada com a lei de Benford, enquanto $\Delta = 4,4$ (como na figura 34) implica que a probabilidade de algum dígito $s \in 1, 2, \ldots, 9$ diferir de $\log(1 + s^{-1})$ é de, no máximo, $4,4\%$.

A distribuição dos dígitos iniciais em cada amostra seguiu satisfatoriamente a lei de Benford. Em relação aos perfis de corrosão referente aos inoxidáveis 304 e 316, os percentuais para um dígito significativo ficaram dentro da margem de um desvio padrão para a média de todas as réplicas utilizadas, com exceção, majoritariamente, do dígito inicial 2. Tal diferença pode estar associada à presença de *outliers* no conjunto, padronização do procedimento por conta do operador, ou mesmo, algum artefato experimental não identificado. De qualquer forma, os resultados mostraram que os dados não apresentam suspeita de terem sido manipulados e que, pelo menos em relação a lei de Benford, são confiáveis segundo esse critério empírico e representam bem as grandezas mensuradas de interesse.

4.2 Mapas de difusão na busca por *outliers* de curva de polarização

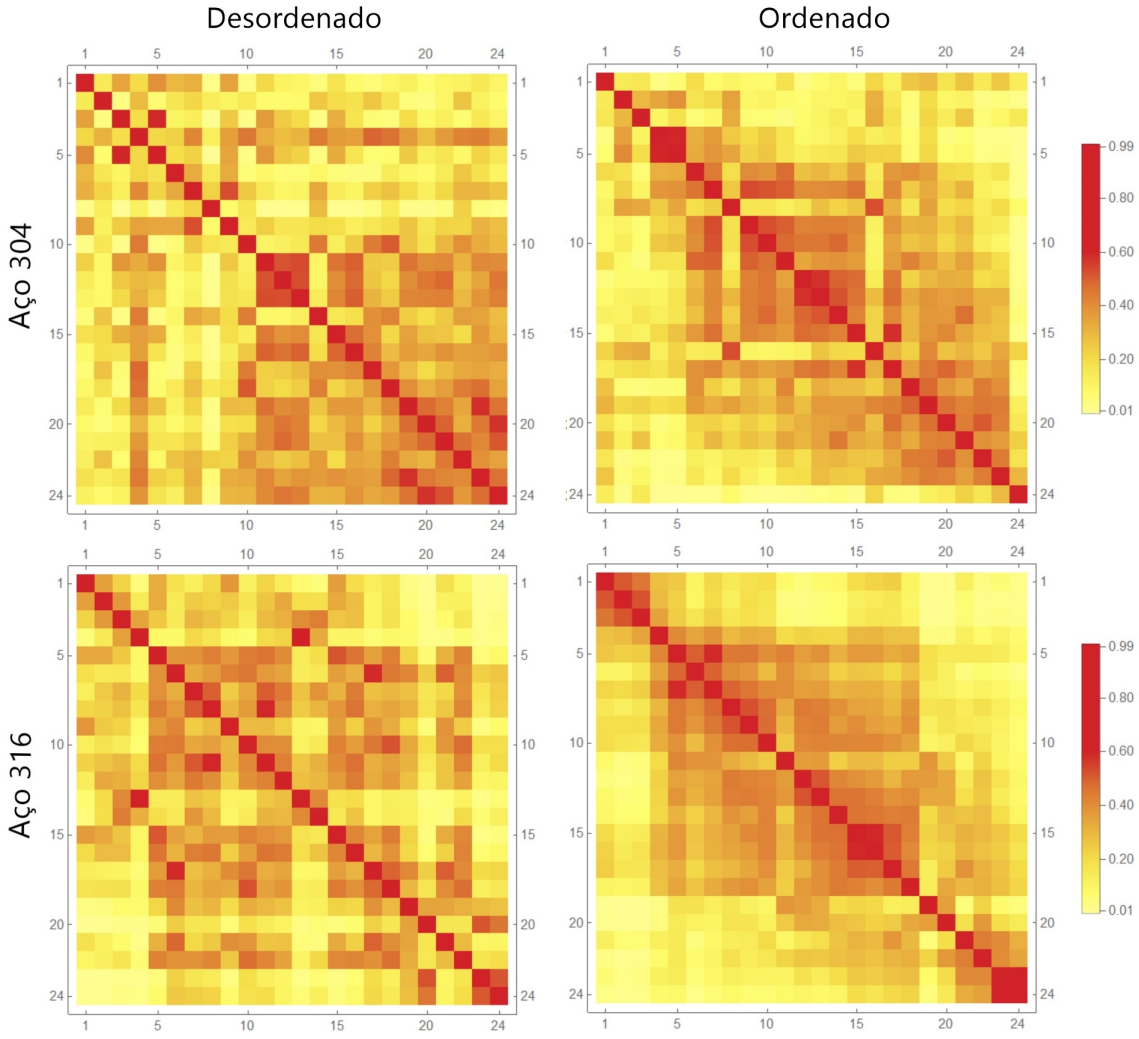
O primeiro passo na busca dos *outliers* da amostra de cada tipo de aço é a execução do algoritmo para cada tipo de aço. Com o objetivo de encontrar os *outliers*, a matriz de difusão P deve ser o foco principal. Tal matriz é responsável por mostrar a semelhança entre os perfis, no sentido de caminhos curtos de alta probabilidade. Para tal, o algoritmo precisa ser executado com um pequeno valor para o parâmetro de escala, o que torna a técnica mais sensível à busca por *outliers*. O limite de conectividade é uma boa sugestão. Os resultados trazidos neste capítulo foram obtidos considerando $\alpha = (\varepsilon_{\mathrm{m}})^2$.

Após obter as matrizes de difusão dos dados referentes a cada um dos aços e os autovetores relacionados a esta matriz, o passo seguinte é ordenar as suas linhas e colunas, segundo o primeiro autovetor não constante. A matriz de difusão ordenada agrupa os perfis, e qualquer bloco (grande o suficiente) pode ser considerado não-outlier. A justificativa da eficiência do uso deste autovetor na reordenação de um dado conjunto é que a ordenação dos perfis definidos pelo mapeamento 1D deve ser tal que perfis semelhantes tenham suas

coordenadas com valores próximos.

A figura 36 mostra como o mapa de difusão unidimensional reorganiza de forma reveladora os perfis exibindo uma estrutura encoberta no conjunto. Assim como realizado anteriormente com a matriz K, a matriz de difusão é vista 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. Esta representação é conhecida como mapa de cores da matriz de difusão.

Figura 36 - Representação das matrizes de difusão dos aços 304 e 316 com $t = 1$.



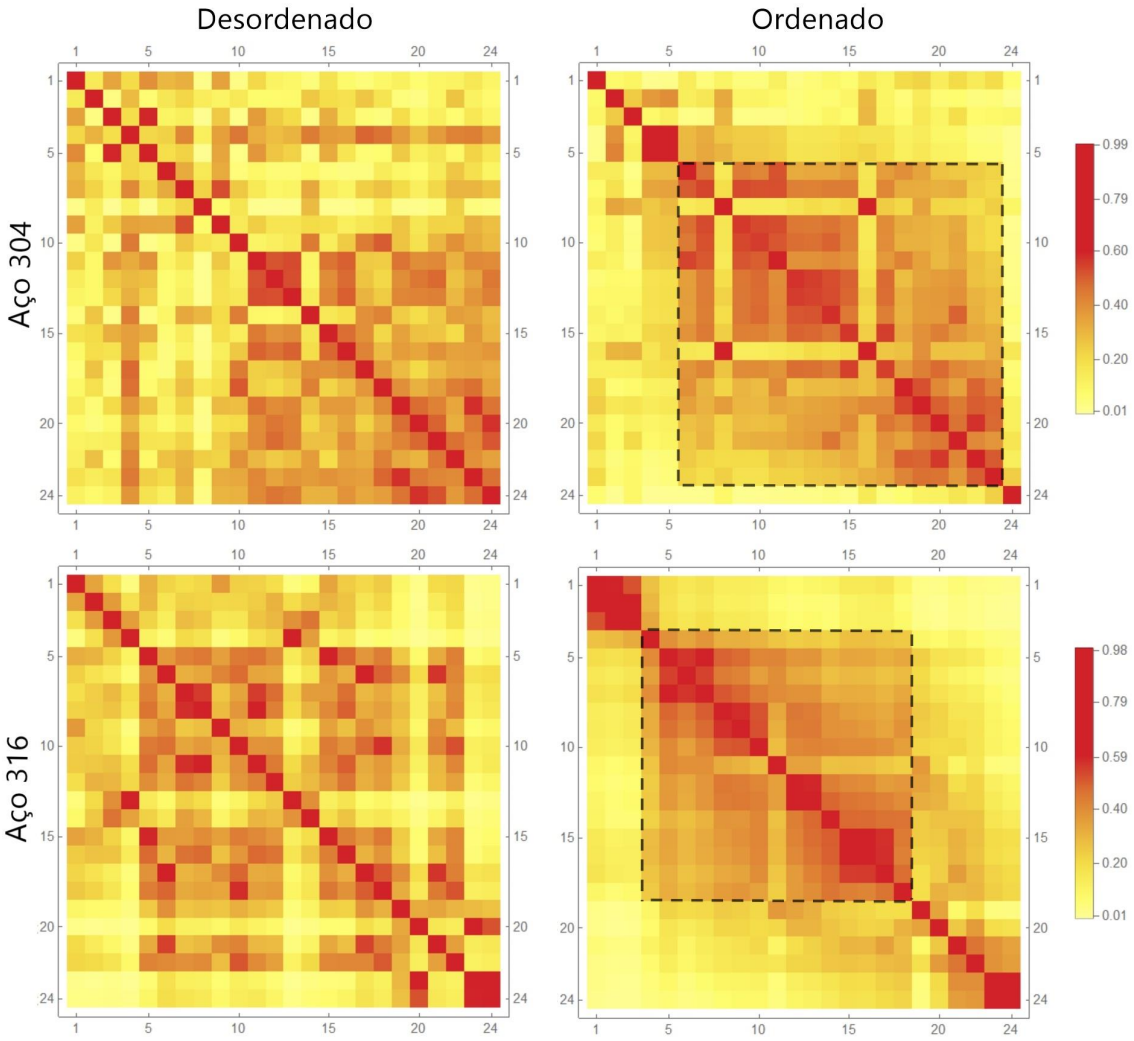
Fonte: O autor, 2022.

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, pode-se aumentar o tempo de difusão t, o que equivale a calcular as potências de P, P^t, possibilitando conectar melhor os perfis pela evolução do

processo de difusão (Fig. 37).

Figura 37 - Mapa de cores dos aços 304 e 316 com $t = 5$.



Fonte: O autor, 2022.

Analisando agora a figura 37, é possível observar com maior clareza as informações obtidas anteriormente. Os blocos aparentes nos dois mapas de cores relativo às matrizes de difusão ordenadas dão o indicativo de 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, *outliers* neste conjunto.

Como mencionado anteriormente, a hipótese para o uso do mapas de cores para a busca de *outliers* é que perfis semelhantes, com características afins em algum sentido, devem ser mapeados em regiões próximas no espaço de difusão. Com isso, qualquer bloco (grande o suficiente) no mapa de cores é considerado não-*outlier*.

Os resultados então podem ser obtidos. Referente ao aço 304, tem-se dois grupos:

$B_{304} = \{6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23\}$, relativo ao maior bloco, e *outliers* $O_{304} = \{1, 2, 3, 4, 5, 24\}$. Em relação ao aço 316, tem-se também 2 grupos: $B_{316} = \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}$, referente ao maior bloco, e *outliers* $O_{316} = \{1, 2, 3, 19, 20, 21, 22, 23, 24\}$. Como observado, seguindo a hipótese anterior, os perfis que não se agrupam dentro do bloco principal são julgados discrepantes.

Devido à ordenação feita pelo primeiro autovetor não constante, os 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 10 tem-se o resultado proposto.

Tabela 10 - 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

Fonte: O autor, 2022.

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.

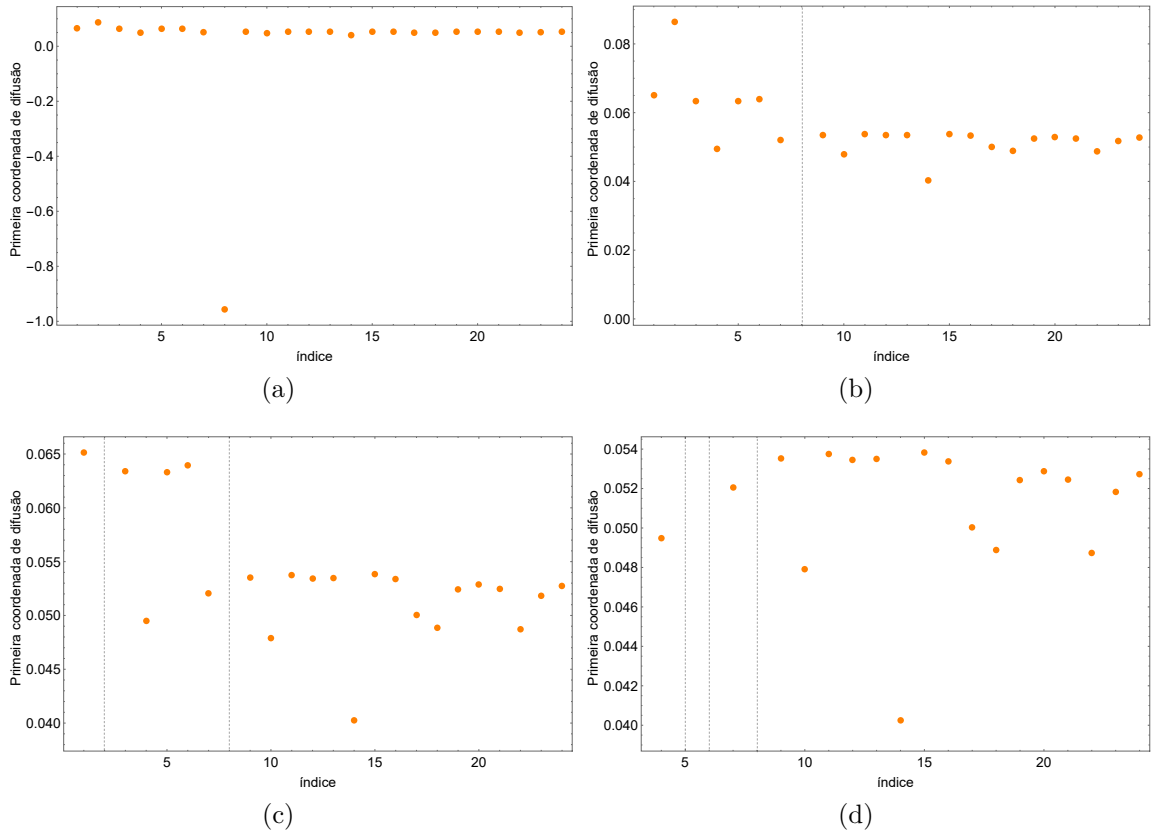
Supondo que para a amostra do aço 304 tenham-se exatamente 6 *outliers*, a probabilidade das três primeiras observações estar entre eles é provavelmente baixa. De igual forma, supondo para o aço 316 exatamente 9 *outliers*, a probabilidade das 4 primeiras observações estarem entre os perfis discrepantes,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

Calculando tais probabilidades, encontrou-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.

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 figuras 38 e 39 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 dos que restaram. Apesar de ser uma investigação subjetiva (que depende da análise visual), o procedimento pode complementar o diagnóstico pelo mapas de cores.

Figura 38 - Mapeamento 1D para as amostras do aço 304. Em cada subfigura exclui-se o(s) perfil(is) mapeado(s) mais distante.

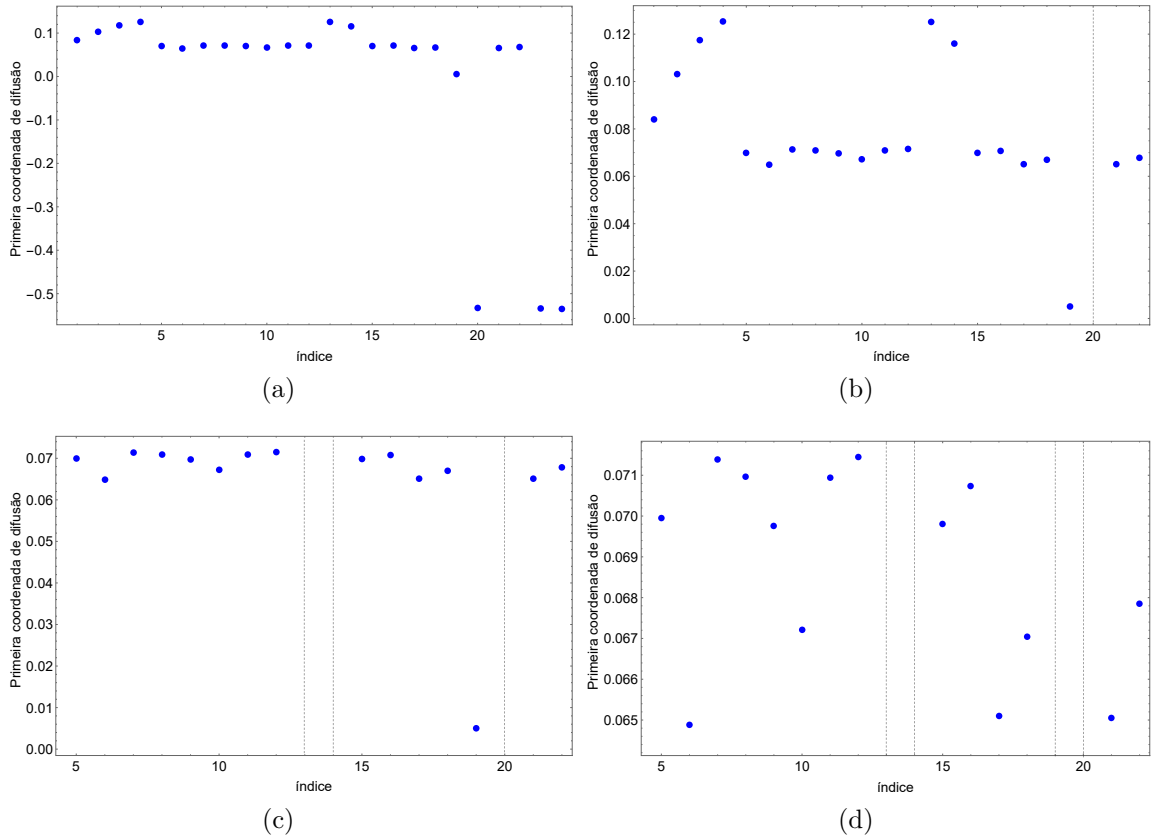


Fonte: O autor, 2022

Indicamos o procedimento mostrado na figura 38. Na subfigura 38a, tem-se o mapeamento 1D com os 24 perfis do aço 304. A seguir, na subfigura 38b, é excluído o perfil de nº 8 por estar mais distante. A subfigura 38c traz agora o mapeamento também sem o perfil nº 2. Por fim, a última subfigura 38d exclui os de números $1, 3, 5$ e 6 (além dos de n° 2 e 8).

Em relação ao aço 316, a subfigura 39a traz o mapeamento 1D com os 24 perfis agora do aço 316. Em 39b, são excluídos os perfis de nº $20, 23$ e 24. A seguir, em 39c, os de números $1, 2, 3, 4, 13$ e 14 e, em seguida, na subfigura 39d, o de nº 19.

Figura 39 - Mapeamento 1D para as amostras do aço 316. As subfiguras mostram o mapeamento a cada extração do(s) possível(is) *outlier*(s).

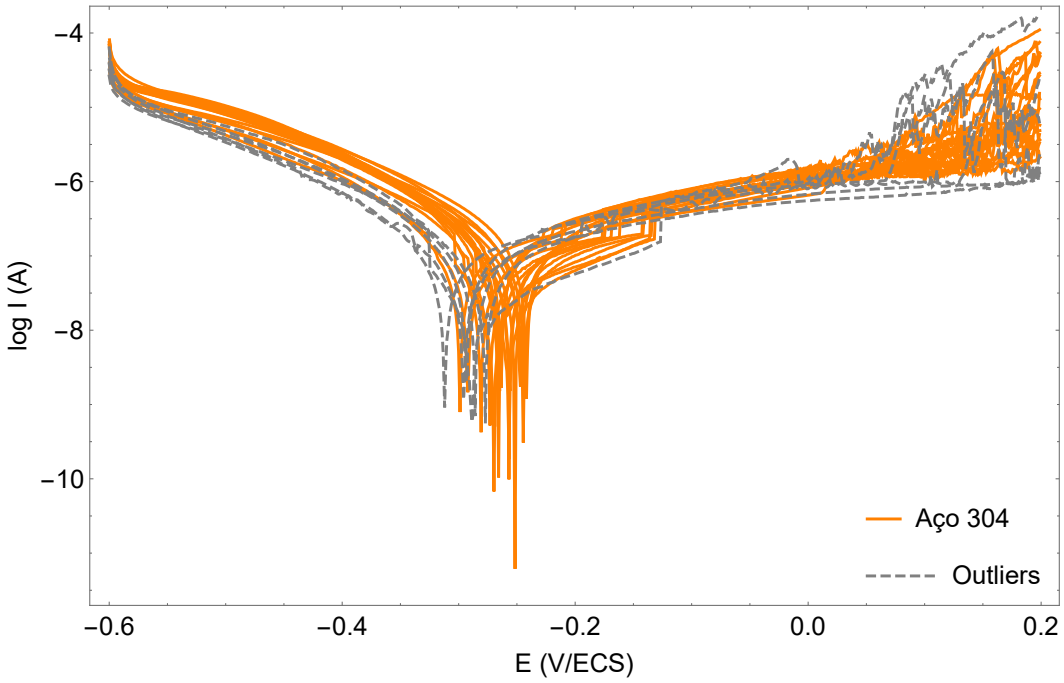


Fonte: O autor, 2022

Fazendo-se 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 figura 38 traz como *outliers* do aço 304 os perfis de nº $1, 2, 3, 5, 6$ e 8. Por outro lado, a figura 39 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.

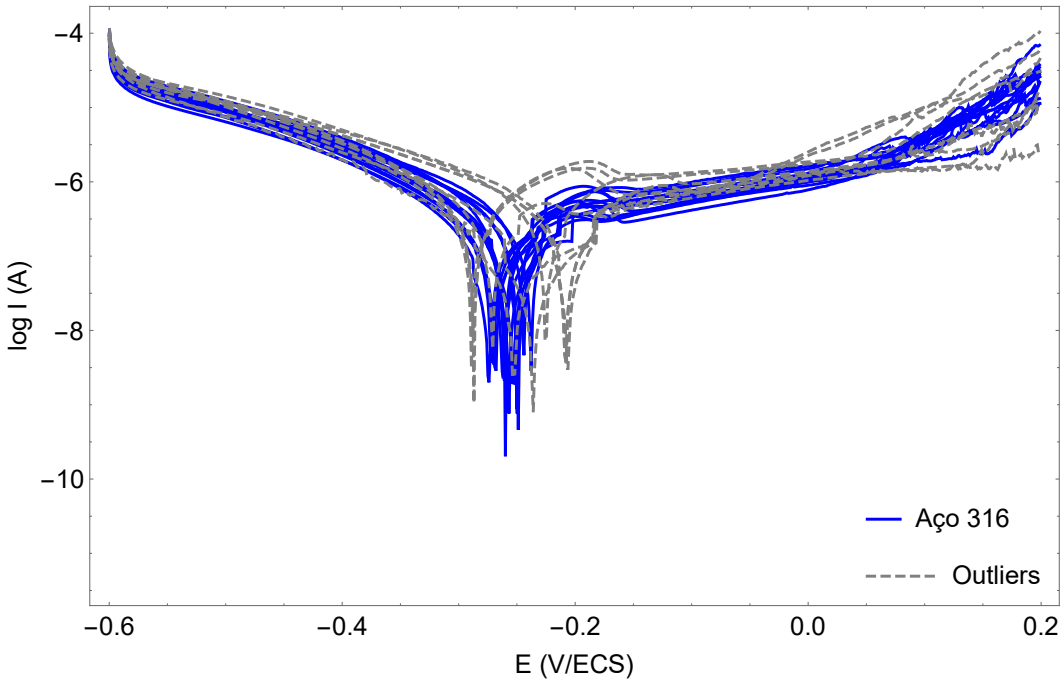
As curvas de polarização, agora sem e com *outliers* de cada aço, são exibidas nas figuras 40 a 42.

Figura 40 - Curvas de polarização experimental do aço 304 com seus possíveis *outliers*.



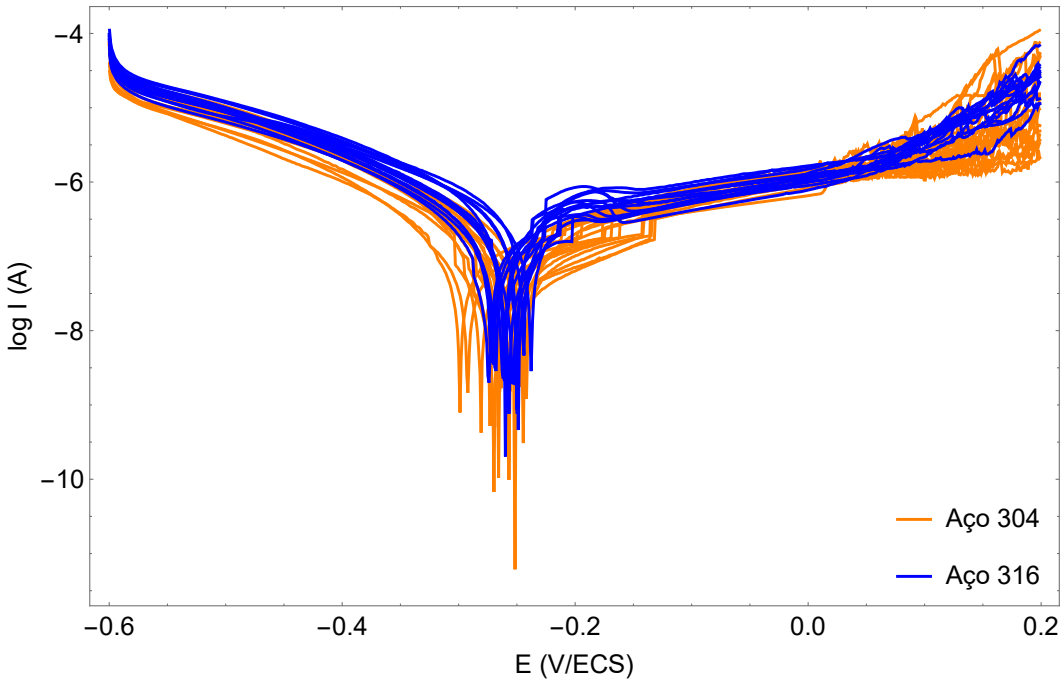
Fonte: O autor, 2022.

Figura 41 - Curvas de polarização experimental do aço 316 com seus possíveis *outliers*.



Fonte: O autor, 2022.

Figura 42 - Curvas de polarização experimental do aços sem os possíveis *outliers*.



Fonte: O autor, 2022.

4.3 ***Outliers versus* não-*outliers***

Identificar *outliers* em um conjuntos de dados experimentais é de fundamental importância na análise de qualquer fenômeno. Entendê-los é essencial em uma análise de dados por pelo menos dois motivos: os *outliers* podem enviesar os resultados, prejudicando toda análise e/ou o seu comportamento pode ser justamente o que está sendo procurado. *Outliers* associados à experimentação podem surgir devido a uma leitura ou erro de anotações e transcrição incorreta dos dados, bem como mudanças não controláveis ou imprevistas nas condições experimentais.

No caso estudado no capítulo anterior, o comportamento dos aços em estudo em relação ao seu potencial de corrosão localizada pode não ter sido bem representado por algumas das medições efetuadas. Se isso ocorreu, os resultados obtidos junto à classificação eficiente podem ter sido influenciados. De fato, isso pode ter várias explicações naturais como erros sistemáticos na fase de realização de aprendizado dos ensaios, problemas de estacionaridade das medidas e/ou dispersão dos parâmetros eletroquímicos.

Após a detecção dos possíveis *outliers* do conjunto de dados dos perfis dos aços como mostrado na seção anterior, o passo seguinte foi reavaliar a classificação dos aços pelo classificador após alcançar a redução de dimensionalidade com os mapas de difusão assim como foi realizado no capítulo 3. A diferença é que agora temos o conjunto de dados mais selecionado e, por hipótese, caracterizando melhor o comportamento dos perfis de

corrosão dos aços em estudo.

De forma análoga, ordenadamente foram realizados as seguintes etapas:

1. Aplicou-se a técnica de mapas de difusão ao conjuntos de perfis considerados não *outliers* dos aços 304 e 316 usando o núcleo gaussiano e experimentando vários parâmetros de escala (α) e parâmetro temporal (t). É importante destacar que agora o conjunto de dados referente aos perfis não *outliers* de ambos os aços entram no algoritmo dos mapas de difusão. Antes, com o intuito de encontrar esses perfis discrepantes em cada amostra dos aços, trabalhou-se com cada grupo separadamente.

2. Obteve-se o mapeamento de acordo com o algoritmo reduzindo a dimensão de 800 para 3. Da mesma forma, o objetivo desta redução é a representação para fins de visualização e maximização do número de informações consideradas.

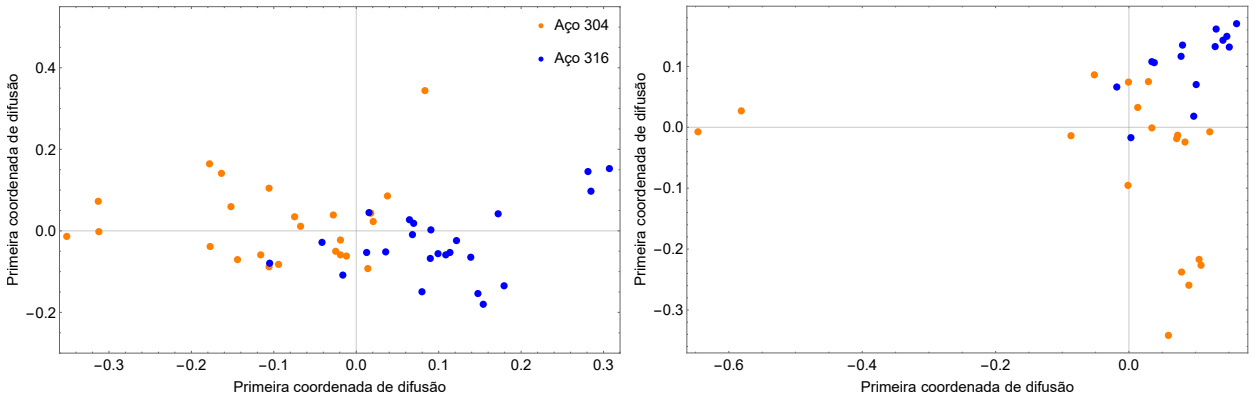
3. Avaliou-se objetivamente o desempenho da abordagem de classificação, usando o método de validação da taxa de acerto e o esquema de validação cruzada de 10 vezes. De forma análoga ao capítulo 3, o conjunto de dados do perfil foi dividido aleatoriamente em 10 dobras, considerando que cada dobra contém dois perfis, um para cada classe do aço. Em cada execução deste esquema, o classificador foi treinado usando todas menos uma dobra e, em seguida, avaliado em como classificou as amostras da dobra separada. A média obtida representa uma medida de desempenho, onde um único número resume a taxa de classificação que representa a proporção geral de sucesso em todas as execuções.

Antes de exibir os resultados obtidos utilizando o classificador *Bayes* nos perfis não-*outliers* e comparar com as taxas de classificação apresentados na tabela 1, é importante comparar também os mapeamentos alcançados com os mapas de difusão agora usando esse seleto grupo em relação ao obtido anteriormente com todos os perfis. As figuras 43 e 44 trazem os mapeamentos 2D com $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $\alpha = (\varepsilon_{\mathrm{d}})^2$, respectivamente, para as amostras dos perfis com e sem *outliers* considerando $t = 2$.

A justificativa para não trazer os mapeamentos para valores menores para o parâmetro α (como $\alpha = (\varepsilon_{\mathrm{m}})^2$ e $\alpha = (\varepsilon_{\mathrm{mean}})^2$, por exemplo) é que, com a depuração dos dados e sua consequente clusterização, valores pequenos para o parâmetro deixam o mapeamento desses dados muito aglomerado dificultando a comparação. Para $\alpha = (\varepsilon_{\mathrm{m}})^2$, por exemplo, os perfis mapeados por meio dos mapas de difusão 2D estão basicamente todos concentrados em torno de uma reta (têm um valor para a primeira coordenada principal bem próximo para todos os dados mapeados—Fig. 45) .

Por outro lado, isso não significa que o classificador teria mais dificuldade. Como o algoritmo trabalha com as coordenadas dos pontos mapeados, ele deve conseguir encontrar

Figura 43 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa completa de potencial.

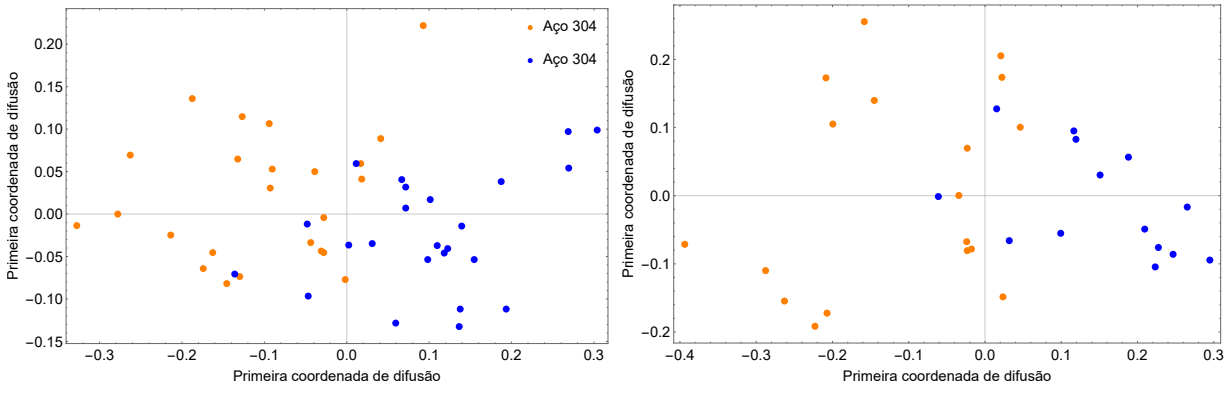


(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não-*outliers*

Fonte: O autor, 2022.

Figura 44 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{d}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa completa de potencial.

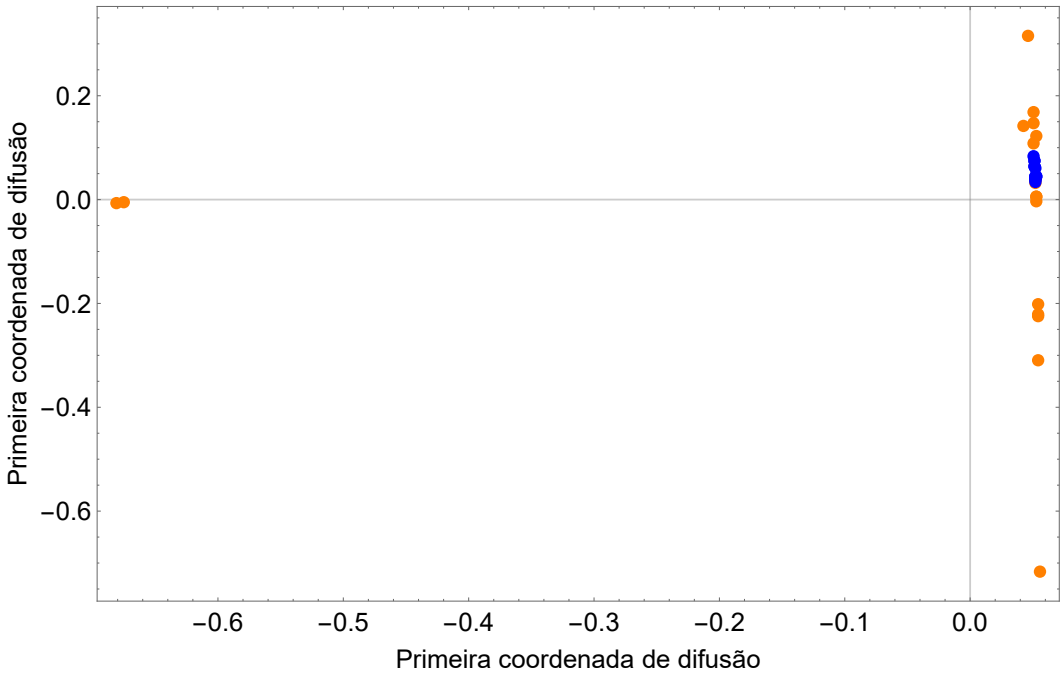


(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não-*outliers*

Fonte: O autor, 2022.

Figura 45 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{m}})^2$ e $t = 2$ para o conjunto dos perfis sem *outliers* para a faixa completa de potencial.



Fonte: O autor, 2022.

o hiperplano[3] separador em alguma escala. A Tabela 11 traz as taxas de classificação usando agora o conjunto de perfis considerados não-*outliers*.

Tabela 11 - Taxas de classificação para o conjunto de perfis não-*outliers* com diferentes escolhas do parâmetro de escala α.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{\mathrm{mean}})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$(\varepsilon_{\mathrm{d}})^2$
#PCC	27	25	26	27
Tc(%)	84	78	81	84

Fonte: O autor, 2022.

Após realizar a leitura dessa tabela e comparar com os resultados obtidos na tabela 1, pode-se, enfim, avaliar o efeito dos *outliers* no conjunto. Nesta análise, pelo menos dois pontos chamam a atenção. O primeiro é que, assim como se imaginava, houve melhorias nas taxas de classificação alcançadas com o classificador em todos os valores utilizados para o parâmetro α, mostrando que a técnica de mapas de difusão é útil na

[3] Hiperplano é a generalização do plano em diferentes números de dimensões. Em particular, num espaço tridimensional um hiperplano é um plano habitual.

busca de *outliers* em um conjunto de dados. A diferença mais significativa ocorreu para os valores do parâmetro com $\alpha = (\varepsilon_{\mathrm{m}})^2$ e $\alpha = (\varepsilon_{\mathrm{mean}})^2$.

Em contrapartida, outro ponto mostra que, apesar de haver a diferença entre as taxas de classificação em cada abordagem (com e sem *outliers*), esta não é muito significativa para os valores maiores do parâmetro de escala mostrando também que a técnica é robusta à presença de *outliers* no conjunto. Ou seja, os mapas de difusão conseguem um mapeamento significativo do conjunto de dados mesmo com dados ruidosos. Isto pode ser visto com valores maiores de α. Para $\alpha = (\varepsilon_{\mathrm{mM}})^2$ ou $\alpha = (\varepsilon_{\mathrm{d}})^2$, por exemplo, as taxas de classificação obtidas com o classificador em cada abordagem foram, basicamente, as mesmas, com diferença de 1 perfil cada (Tab. 12).

Tabela 12 - Resumo comparativo das taxas de classificação obtidas com o classificador *Bayes* para o mapeamento com e sem *outliers*.

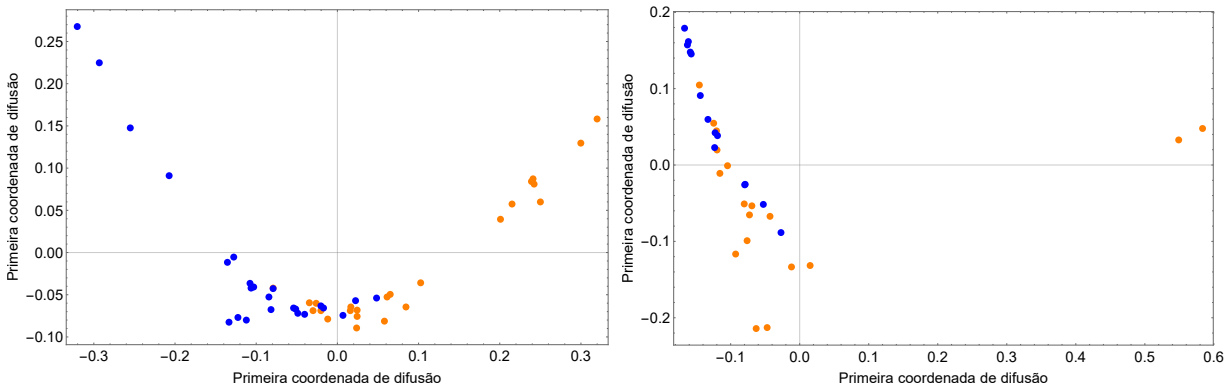
	Tc (%)	
$\boldsymbol{\alpha}$	Com *outliers*	Sem *outliers*
$(\varepsilon_{\mathrm{m}})^2$	54	84
$(\varepsilon_{\mathrm{mean}})^2$	63	78
$(\varepsilon_{\mathrm{mM}})^2$	79	81
$(\varepsilon_{\mathrm{d}})^2$	81	84

Fonte: O autor, 2022.

4.3.1 Faixa de baixo e alto potencial

Da mesma forma como anteriormente, o interesse aqui é apresentar e avaliar o ganho obtido com a depuração dos dados por meio da técnica de mapas de difusão agora considerando somente as faixas de baixo e alto potenciais. As figuras 46 e 47 exibem os mapeamentos alcançados com os mapas de difusão 2D para $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $\alpha = (\varepsilon_{\mathrm{d}})^2$, respectivamente, para as amostras dos perfis com e sem *outliers*, com $t = 2$, na faixa de baixo potencial. As figuras 48 e 49 fazem o mesmo para a faixa de alto potencial. A tabela 13 traz os resultados de classificação para os perfis selecionados para a faixa de baixo potencial nos dois aços e a tabela 14 traz os resultados de classificação para os perfis selecionados para a faixa de alto potencial. Por fim, a tabela 15 compara os resultados obtidos com as duas abordagens (com e sem *outliers*) para ambas as faixas de potencial.

Figura 46 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa de baixo potencial.

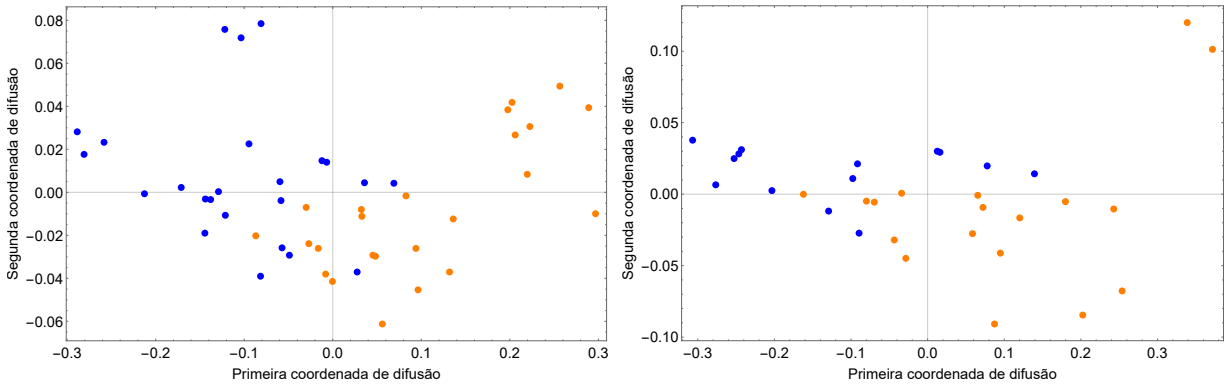


(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não-*outliers*

Fonte: O autor, 2022.

Figura 47 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{d}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa de baixo potencial.



(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não-*outliers*

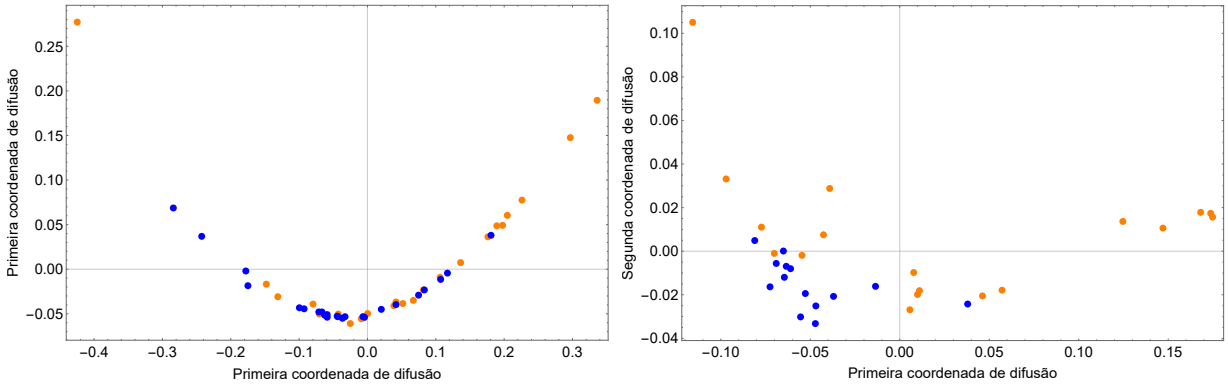
Fonte: O autor, 2022.

Tabela 13 - Taxas de classificação (Tc) para o conjunto de perfis não *outliers* na faixa de baixo potencial para diferentes escolhas dos parâmetros α.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{\mathrm{mean}})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$(\varepsilon_{\mathrm{d}})^2$
#PCC	17	23	26	26
Tc(%)	53	72	81	81

Fonte: O autor, 2022.

Figura 48 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{mM}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa de alto potencial.

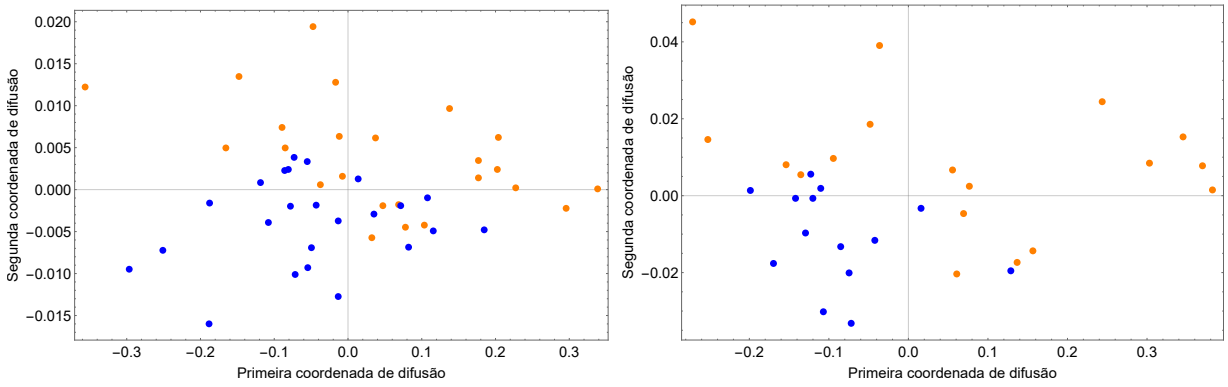


(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não-*outliers*

Fonte: O autor, 2022.

Figura 49 - Mapeamento 2D com $\alpha = (\varepsilon_{\mathrm{d}})^2$ e $t = 2$ para o conjunto dos perfis com e sem *outliers* para a faixa de alto potencial.



(a) Mapeamento 2D de todos os 48 perfis

(b) Mapeamento 2D dos 32 perfis considerados não *outliers*

Fonte: O autor, 2022.

Tabela 14 - Taxas de classificação (Tc) para o conjunto de perfis não-*outliers* na faixa de alto potencial para diferentes escolhas dos parâmetros α.

$\boldsymbol{\alpha}$	$(\varepsilon_{\mathrm{m}})^2$	$(\varepsilon_{\mathrm{mean}})^2$	$(\varepsilon_{\mathrm{mM}})^2$	$(\varepsilon_{\mathrm{d}})^2$
#PCC	25	26	27	30
Tc(%)	78	81	84	94

Fonte: O autor, 2022.

Tabela 15 - Resumo das taxas de classificação (Tc) para o conjunto de perfis com e sem *outliers* nas faixas de baixo e alto potencial para diferentes escolhas dos parâmetros α.

$\boldsymbol{\alpha}$	**Taxa de classificação (%)**			
	Baixo potencial		**Alto potencial**	
	Com *outliers*	Sem *outliers*	Com *outliers*	Sem *outliers*
$(\varepsilon_{\mathrm{m}})^2$	44	53	56	78
$(\varepsilon_{\mathrm{mean}})^2$	60	72	67	81
$(\varepsilon_{\mathrm{mM}})^2$	79	81	60	84
$(\varepsilon_{\mathrm{d}})^2$	88	81	79	94

Fonte: O autor, 2022.

Levando-se em conta o que é registrado na tabela 15, é possível também concluir o efeito positivo da depuração dos dados para as diferentes faixas por meio da técnica de mapas de difusão. Em ambas as parcelas, as taxas de classificação sofreram aumentos significativos, principalmente na faixa de interesse de alto potencial. No melhor resultado obtido com o parâmetro $(\alpha = \varepsilon_{\mathrm{d}})^2$, a taxa de acertos do classificador passa de 79% para 94%. Isso evidencia, primeiramente, que os mapas de difusão são úteis na busca de *outliers* em um conjunto de dados e que a presença desses na amostra prejudica de forma diferente a classificação eficiente dos perfis de acordo com a faixa de potencial analisada.

4.4 Mapas de difusão na busca por *outliers* em sinais de impedância eletroquímica

Assim como os mapas de difusão foram úteis na busca de *outliers* nos perfis das curvas de polarização dos aços 304 e 316 e possibilitaram definir um perfil representativo para todas as amostras, a rotina também foi aplicada a perfis de resposta provenientes da técnica de espectroscopia eletroquímica. Tal procedimento, conhecido simplesmente como impedância eletroquímica, consiste na aplicação de um potencial elétrico variável, por meio de um potenciostato em uma célula eletroquímica e a medição da resposta desta célula a esta excitação na forma de corrente elétrica. Cabe a essa seção apresentar os resultados da aplicação dos mapas de difusão na busca de *outliers* também para esse tipo de perfil.

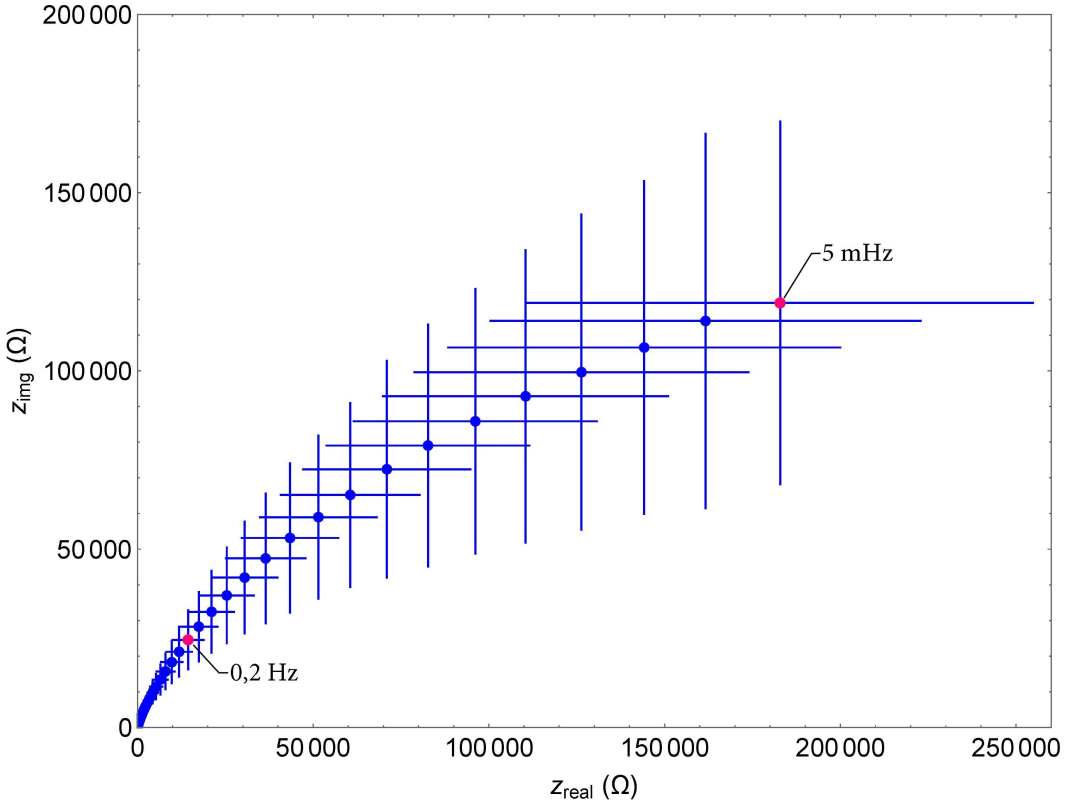
4.4.1 Contexto experimental

O material usado nos experimentos de impedância eletroquímica foi o aço inoxidável superdúplex UNS S32750. Como feito anteriormente, usa-se agora a abreviação 327. A faixa de frequência empregada nos ensaios foi de 20 kHz a 5 mHz, e amplitude de 10 mV de valor eficaz. A impedância eletroquímica é um número complexo cujas partes reais e imaginárias variam com a frequência de excitação. Quando a impedância é apresentada na forma da parte imaginária *versus* real, denomina-se forma de Nyquist; e quando é graficada na forma de log módulo *versus* log da frequência e fase vs. log da frequência é chamada de forma de Bode. Estes modos de apresentação, entretanto, são formas distintas do mesmo espectro de impedância. O módulo da impedância é expresso em Ω, e não $\Omega \cdot \mathrm{cm}^2$. Estas correções para as áreas, ainda que fundamentais em eletroquímica, não influem na análise realizada.

Na figura 50 são apresentados todos os diagramas de impedância eletroquímica do aço 327, na forma de Nyquist, isto é, parte imaginária vs. real, para cada frequência. Ao todo, 30 diagramas foram usados neste trabalho e, para cada variável medida, 69 registros foram feitos para diferentes frequências. Nota-se que para cada frequência há uma dispersão das duas componentes, cujos desvios padrões são dados pelas barras. Próximo da origem dos eixos a frequência é elevada, e à medida que os pontos se afastam da origem a frequência diminui até cerca de 5 mHz.

Assim como realizado anteriormente com os perfis de corrosão, cujo objetivo foi a busca de *outliers* na amostra, a matriz de difusão P deve ser o foco principal. É ela que, após a organização segundo o primeiro autovetor não constante, fornece o bloco de perfis afins segundo a distância de difusão e exibe os que destes distanciam. No intuito de fazer uma análise mais aprofundada sobre os perfis apresentados, vários testes foram realizados. Foi analisado a similaridade primeiramente entre grupos que compartilham o mesmo comportamento em relação ao log do módulo, à fase, e, por fim, agrupando-se as duas informações juntas ao algoritmo. É observado que os resultados se complementam e que, com isso, a técnica também pode ser utilizada para este fim.

Figura 50 - Diagrama de impedância médio e o desvio padrão das componentes imaginária e real do aço inoxidável austenítico 32750.



Fonte: O autor, 2022.

4.4.2 Sinais de impedância e a lei de Benford

Antes de exibir os resultados propostos com os *outliers*, assim como com as curvas de polarização, esta subseção traz a aplicação da lei de Benford como uma etapa de pré-processamento para os dados de impedância. Como visto anteriormente, o procedimento pode ser útil para verificar possíveis problema com a qualidade dos dados obtidos e garantir um atestado que o experimento não foi realizado em condições duvidosas. Uma desconformidade com a lei, no entanto, pode ser atribuída a uma multitude de razões.

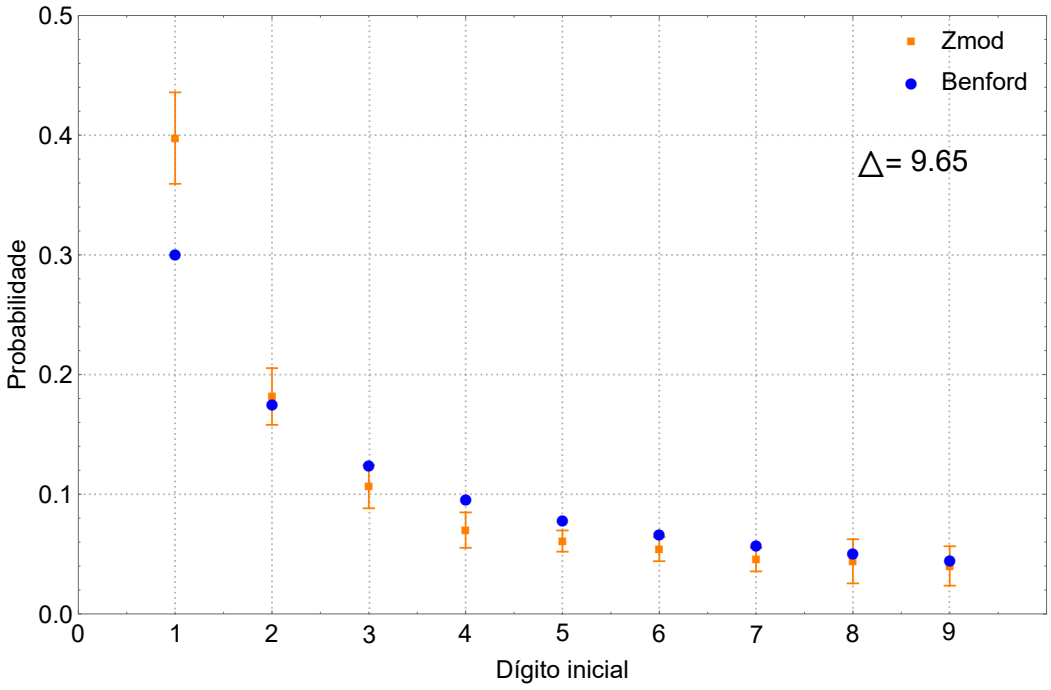
Com os dados referentes às curvas de impedância, agora a aplicação envolveu 30 diagramas onde foram consideradas as informações sobre o módulo de cada sinal. Cada um destes, representados aqui por $|z_i|$, com $i \in \{1, 2, \ldots, 69\}$, consiste das respostas obtidas em relação às partes real e imaginária dos complexos correspondentes segundo a equação 35. Tais medidas definem o módulo de cada sinal e variam de acordo com a frequência a

que cada amostra é submetida.

$$|z_i(f)| = \sqrt{(z\text{real}_i(f))^2 + (z\text{imag}_i(f))^2} \tag{35}$$

Ao aplicar a lei dos dígitos significativos aos módulos de cada perfil, obteve-se a média do grupo e seu desvio-padrão. As figuras 51 e 52 mostram, respectivamente, a distribuição do primeiro e segundo dígitos significativos do módulo da impedância em comparação com a distribuição teórica de Benford.

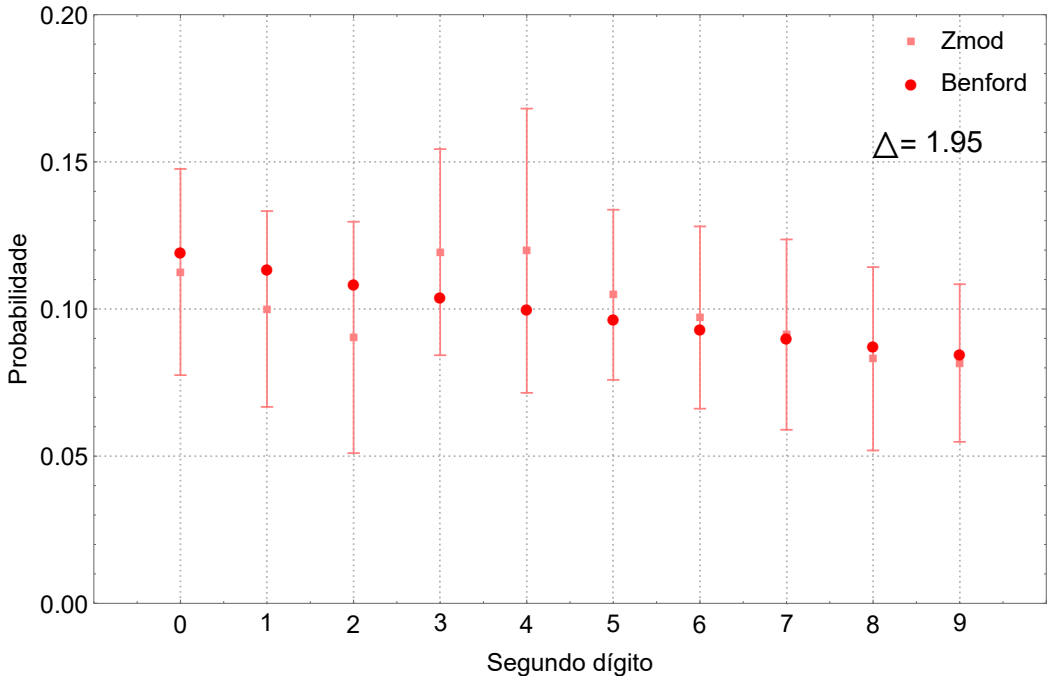
Figura 51 - Distribuição média dos primeiros dígitos dos 30 perfis do aço 327 com seu desvio-padrão e a distribuição segundo Benford do módulo dos sinais em valores absolutos.



Fonte: O autor, 2022.

Com os dados de impedância referentes à amostra do aço 327, novamente a curiosa distribuição apareceu. Para o primeiro dígito, apesar da frequência dos números $1, 4$ e 5 ficarem fora da margem de um desvio padrão das amostras, o coeficiente Δ com valor de $9,65$ está diretamente associado ao dígito 1. O fato de a probabilidade deste dígito ser acima daquela prevista por Benford pode ser causada pelo fato da faixa de frequência (em Hz) do ensaio experimental ser sempre finita, no presente caso de 5 mHz a 20 kHz. Possivelmente, quanto maior a faixa de frequências utilizada, mais a distribuição se aproxime da previsão teórica.

Figura 52 - Distribuição média dos segundos dígitos dos 30 perfis do aço 327 com seu desvio-padrão e a distribuição segundo Benford do módulo dos sinais em valores absolutos.



Fonte: O autor, 2022.

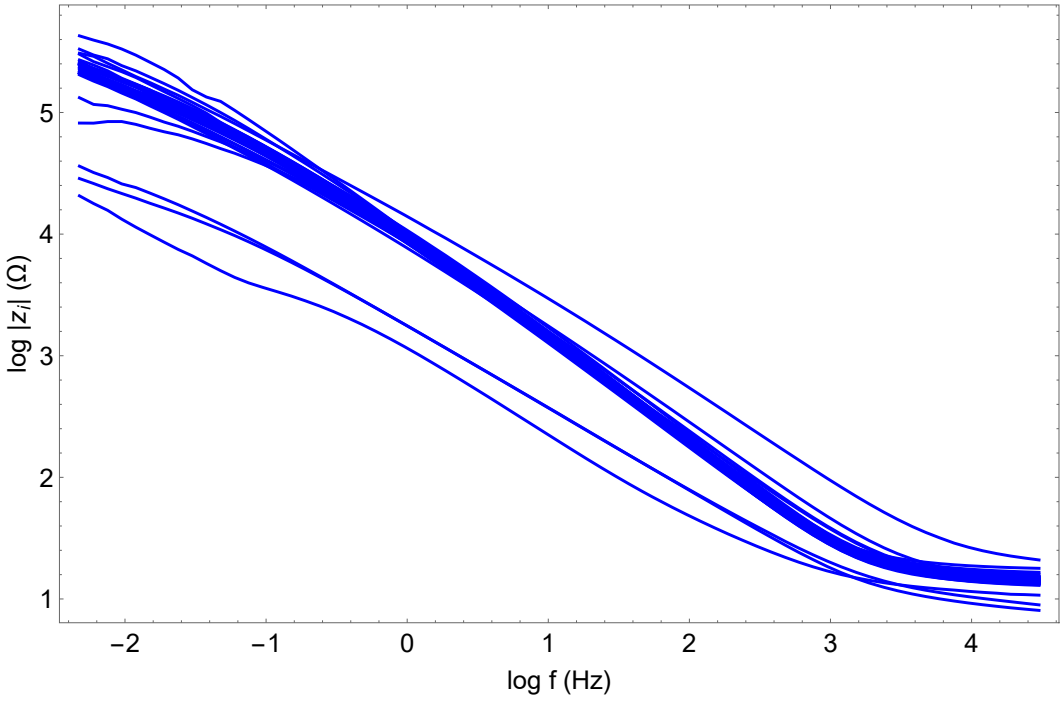
Para o segundo dígito, obteve-se um resultado ainda melhor. Com $\Delta = 1,95$, a probabilidade de algum dígito diferir da distribuição teórica de Benford para o segundo dígito é de, no máximo, $1,95\%$, garantindo a conformidade com a distribuição. Mais uma vez a lei de Benford se mostrou presente ditando regularidade da frequência dos dígitos significativos.

4.4.3 Mapa de cores e os resultados propostos

4.4.3.1 Logaritmo do módulo *versus* logaritmo da frequência

Na abordagem aqui apresentada, os primeiros resultados trazem somente a informação referente ao módulo de cada sinal de impedância. As curvas que representam o log de $|z_i|$ se agrupam em grande maioria próximo ao centro da figura 53 e é possível também perceber pelo menos 4 curvas com comportamento discrepante. Três delas similares entre si e a quarta mais próxima do grupo com mais amostras. Entretanto, esta análise visual pode não ser confiável cabendo aos mapas de difusão a tarefa de encontrar os perfis *outliers* na amostra.

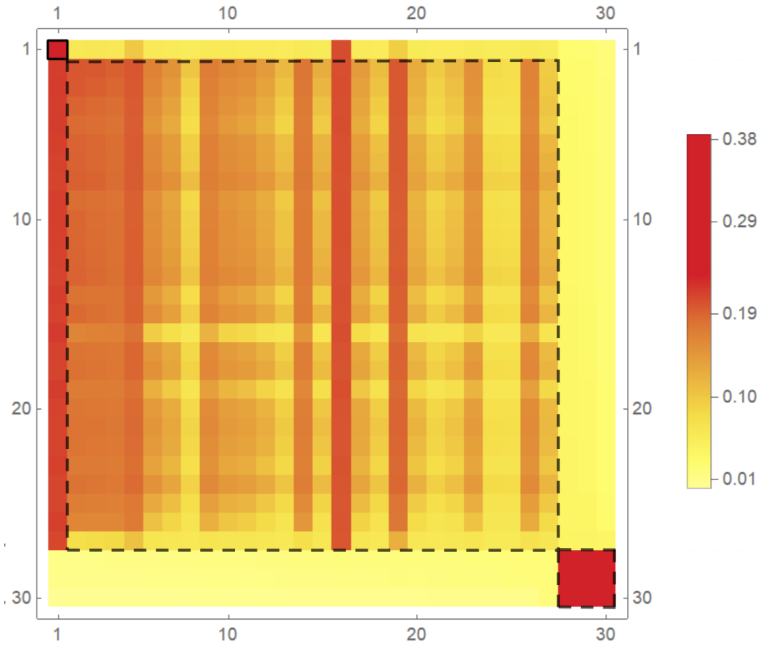
Figura 53 - Forma de Bode com log $|z_i|$ *versus* log f para os 30 perfis de impedância.



Fonte: O autor, 2022.

O mapa de cores exibidos na figura 54 fornece a confirmação de medidas discrepantes. Os blocos de similaridade, agora, exibem um primeiro grupo cujos índices vão de 2 a 26, e dois outros compreendendo o perfil de índice ordenado 1 e outro que compreende os perfis de índices 27 a 29. Apesar destes blocos não serem tão nítidos como os das curvas de polarização visto na seção anterior, é clara a distinção de três grupos aqui denominados A_1, A_2 e A_3. Os quadrados pontilhados na figura 54 destacam tais grupos.

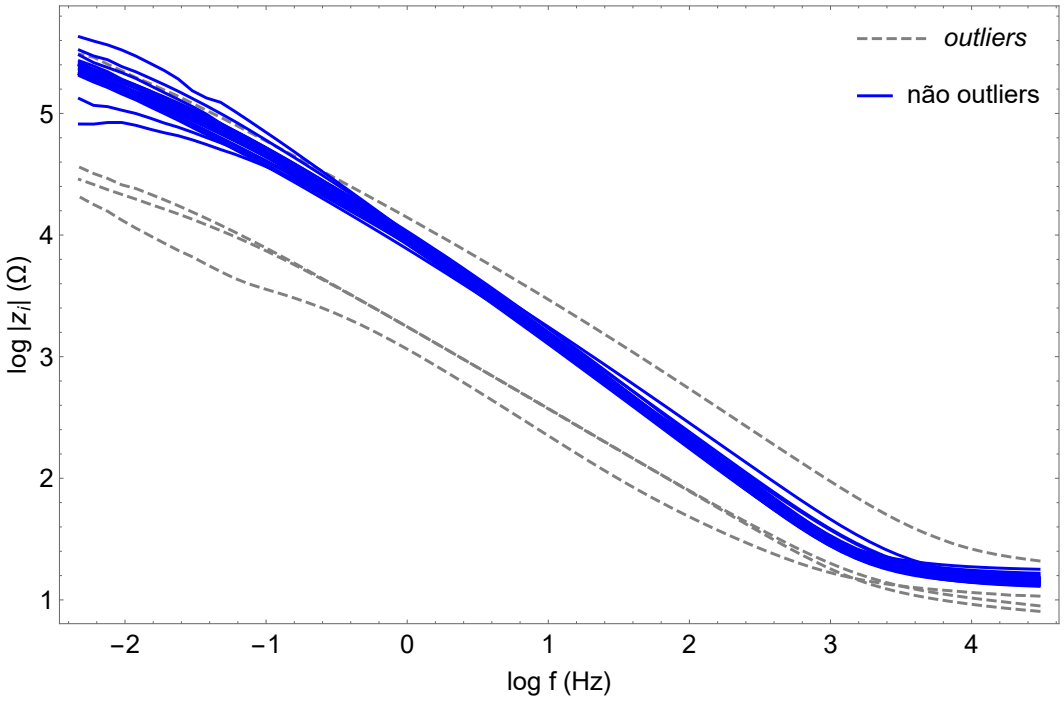
Figura 54 - Mapa de cores ordenado da variável log $|z_i|$ dos perfis de impedância.



Fonte: O autor, 2022.

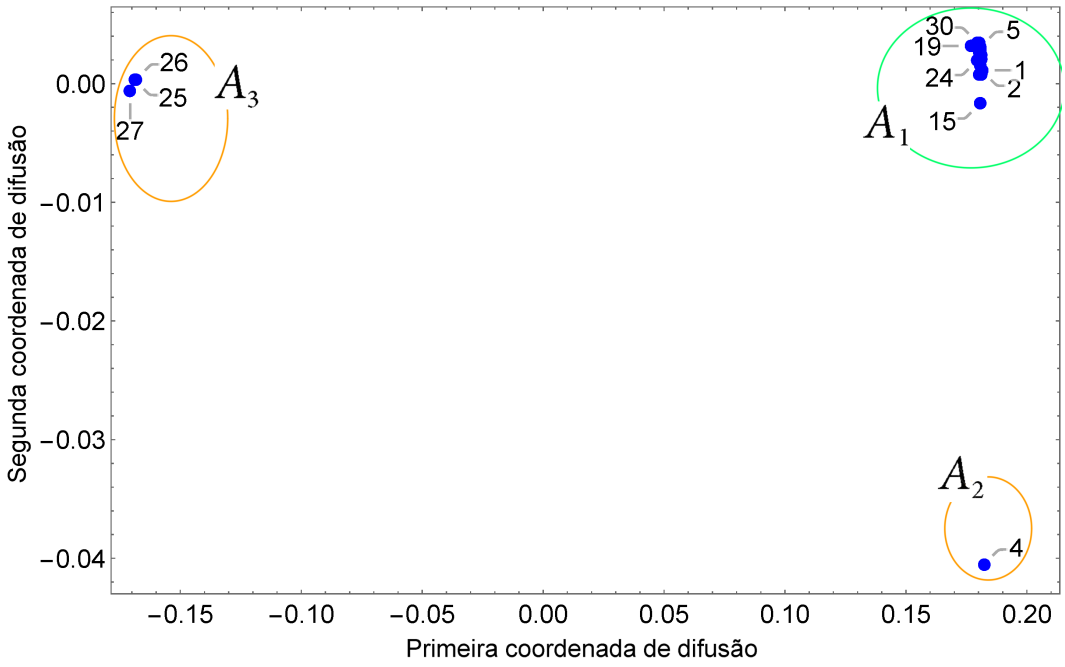
Ao fazer a correspondência correta entre os índices dos perfis ordenados e os índices das amostras, tem-se o resultado proposto. O grupo A_1 é considerado representativo e compreende os perfis de índices 1 a 3, 5 a 24 e 28 a 30. Assim sendo, os grupos A_2 e A_3 compõem os perfis de índices $4, 25, 26$ e 27 e, estes, são considerados *outliers* para a variável em análise. A figura 55 traz agora as curvas destacando os grupos considerados representativos e discrepantes e a figura 56 mostra o mapeamento 2D obtido com os mapas de difusão para a amostra. Como é possível constatar, ambas as figuras exibem os diferentes grupos observados com seus respectivos rótulos.

Figura 55 - *Outliers* e não *outliers* das curvas log $|z_i|$ *versus* log f.



Fonte: O autor, 2022.

Figura 56 - Mapeamento 2D para a variável log $|z_i|$ dos perfis de impedância. Os grupos ficam claramente distintos com o mapeamento.

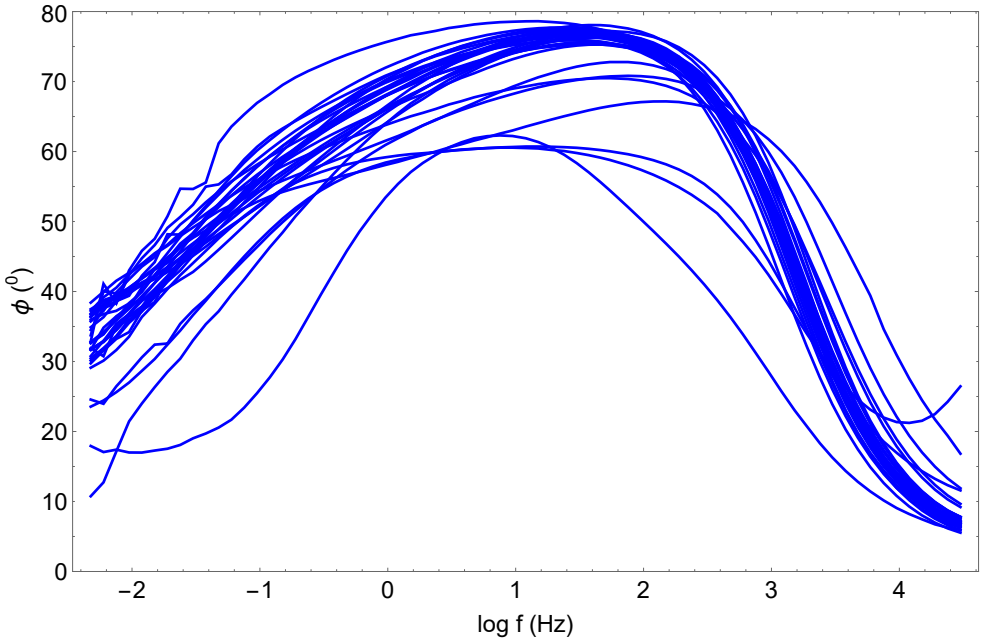


Fonte: O autor, 2022.

4.4.3.2 Fase *versus* log da frequência

Nesta subseção são apresentados os resultados envolvendo somente a informação de fase de cada sinal de impedância. Como cada sinal é representado por um número complexo, a fase aqui é o argumento (ou ângulo) de cada número e representa seus deslocamento angular. A figura 57 exibe os 30 diagramas utilizados nesse trabalho, cujos ângulos de fase variam de 0 a 90°.

Figura 57 - Forma de Bode com ângulo de fase vs. log da frequência para os 30 perfis de impedância.



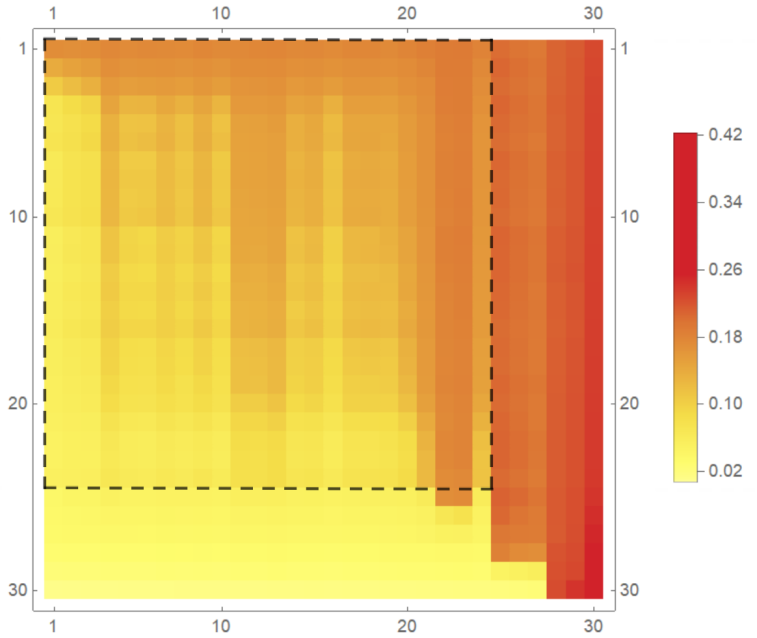
Fonte: O autor, 2022.

Examinando novamente de forma visual, percebe-se agora um grupo de perfis discrepantes com maior número que o anterior. Enquanto que com o diagrama de módulo conseguiu-se imaginar, a princípio, 4 *outliers*, confirmados posteriormente com o mapa de cores, o diagrama com ângulo de fase vs. log da frequência parece exibir mais perfis de comportamento diferente da maioria. Uma inspeção visual poderia indicar de 5 a 8 *outliers*. É esta confirmação que se busca alcançar com a técnica.

O mapa de cores na figura 58 obtido por meio da matriz de difusão ordenada para o conjunto dá o suporte para os resultados apresentados em seguida. O bloco de similaridade, agora, não é tão claro. O quadrado tracejado na figura demarca o único bloco de perfis mais afins com índices que vão de 1 a 24. Os demais, entretanto, não estão tão distantes. Fazendo a correspondência para o índice original, o bloco, intitulado B_1 compreende os perfis de números 1 a 3, 5 a 18, 20 a 24 e 28 a 30. Aqueles que não estão no grupo (B_2) são considerados *outliers*: $4, 15, 19, 25, 26$ e 27. Assim como com as curvas referente ao

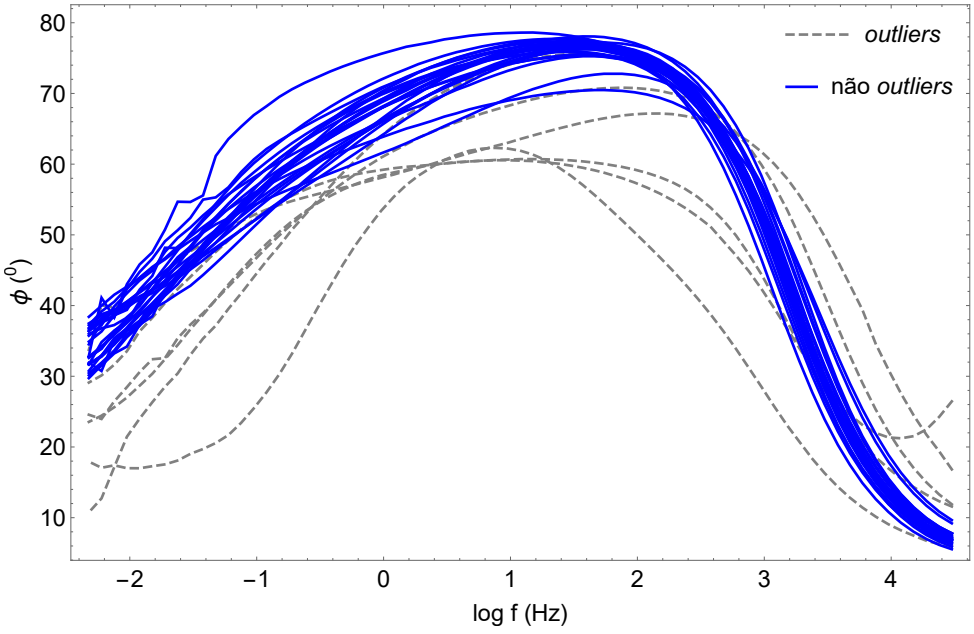
módulo, a figura 59 exibe as curvas com os grupos representativos e discrepantes e a figura 60 traz o mapeamento 2D.

Figura 58 - Mapa de cores ordenado da variável fase dos perfis de impedância.



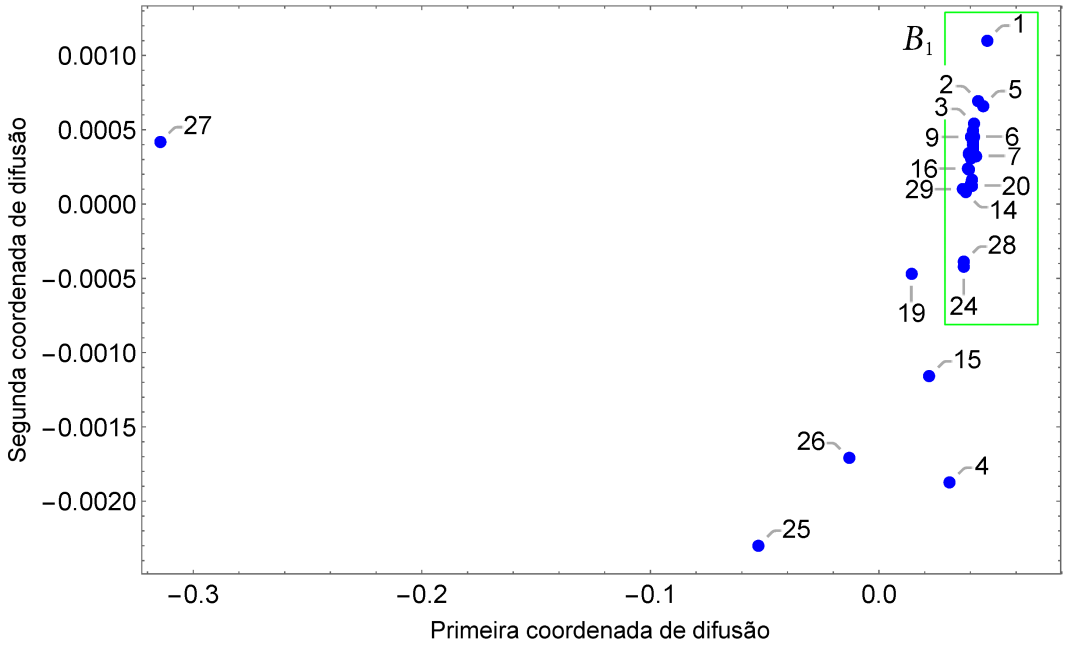
Fonte: O autor, 2022.

Figura 59 - *Outliers* e não-*outliers* das curvas fase vs. log frequência.



Fonte: O autor, 2022.

Figura 60 - Mapeamento 2D para a variável fase dos perfis de impedância.



Fonte: O autor, 2022.

4.4.3.3 Log do módulo e fase *versus* log da frequência

Cabe agora nesta seção fazer a análise de *outliers* considerando as duas informações juntas inseridas no algoritmo. De fato, ainda que possam ser analisadas de forma separada, as variáveis módulo e fase dos perfis de impedância são informações que constituem o sinal e que, naturalmente, se relacionam.

Buscando inserir ambas as informações das variáveis no algoritmo, a primeira mudança em relação à implementação da técnica como foi antes realizada é a definição da norma para as entradas da matriz de similaridade $\tilde{K}$. Como descrito pela equação 10, cada entrada dessa matriz de similaridade é obtida pelo núcleo gaussiano que utiliza a distância euclidiana entre cada ponto de dados visto no espaço de alta dimensão. A diferença agora é que tendo duas informações (módulo e fase), precisamos definir como calcular a norma de uma matriz que, no caso, tem 69 linhas e 2 colunas.

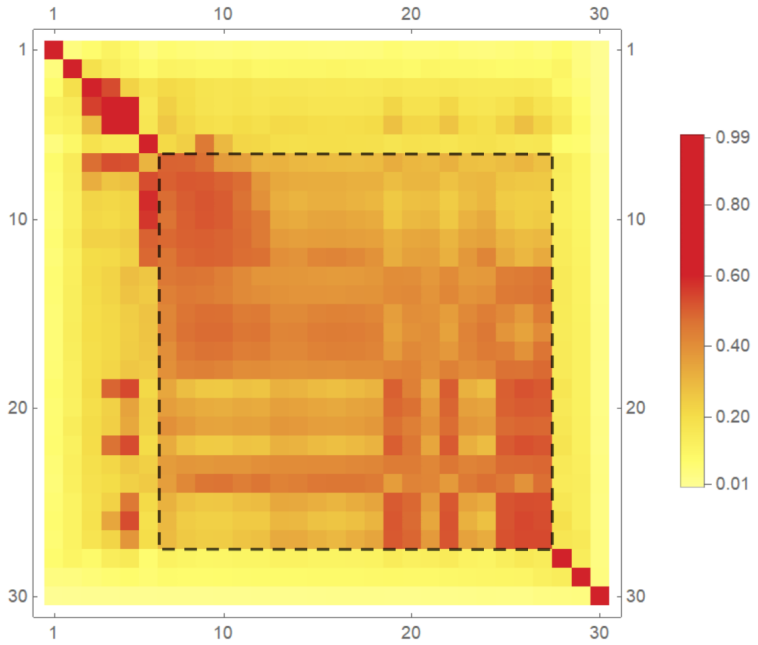
Neste sentido, foi utilizado primeiramente uma normalização das entradas para equilibrar os pesos de cada variável: foi estimada a razão entre o valor numérico em cada entrada e o maior valor registrado nesse perfil para cada um dos perfis. O procedimento é necessário para balancear as entradas e permitir que cada variável contribua de forma equânime com a matriz de similaridade. Em seguida, utilizou-se da norma da matriz de dados $X_{(69\times 2)}$ induzida pela norma vetorial:

$$\|X\| = \max_{\|\mathbf{n}\|=1} \|X\mathbf{n}\| \tag{36}$$

Feito isto, procedeu-se com o algoritmo como descrito no capítulo 2 para encontrar a matriz de difusão e, ordenada pelo maior autovetor não constante, obter o mapa de cores.

A figura 61 traz o mapa de cores. O bloco de similaridades agora, usando também $t = 3$, exibe um bloco de perfis mais afins ao centro com índices que vão de 7 a 27. O tracejado na figura ajuda a guiar os olhos e exibe o único bloco de perfis semelhantes aparentes na amostra. A correspondência para o índice original traz o bloco intitulado C_1 compreendendo os perfis de números 2 a 30, excluindo o 4, o 15, o 19 e os perfis de 24 a 28. Como antes, os que não se encontram no grupo (C_2) são considerados *outliers*: $1, 4, 15, 19, 24, 25, 26, 27$ e 28. Observa-se que com a implementação das duas variáveis juntas no algoritmo, mais perfis foram identificados como discrepantes no conjunto.

Figura 61 - Mapa de cores ordenado das variáveis módulo e fase dos perfis de impedância.



Fonte: O autor, 2022.

4.4.3.4 Resultados da aplicação dos mapas de difusão

A tabela 16 resume os resultados obtidos com a aplicação dos mapas de difusão em cada uma das abordagens: log do módulo, à fase e as duas informações juntas ao algoritmo. Comparando-se cada resultado, verifica-se concordância entre os perfis tomados como discrepantes a cada rotina com alguns representantes apontados sempre em cada abordagem. Isso mostra que a técnica é coerente. É possível também observar que o conjunto de perfis considerados *outliers* com a análise isolada do módulo e da fase é

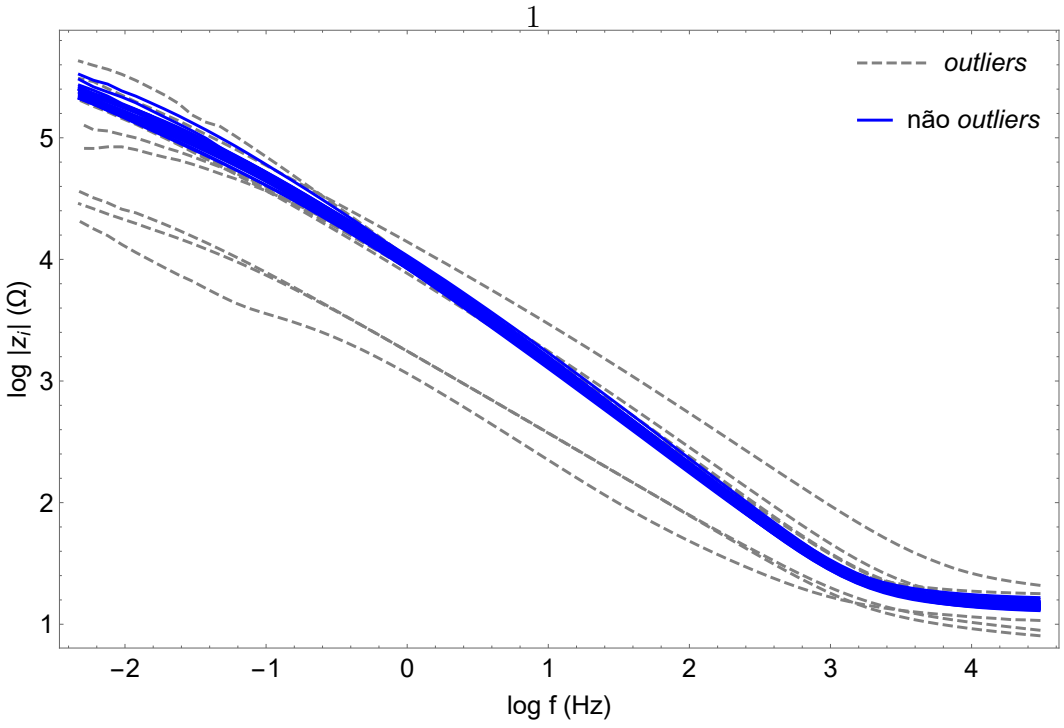
subconjunto dos que foram considerados discrepantes usando as duas informações juntas $((A_2 \cup A_3) \subset C_2$ e $B_2 \subset C_2$. Assim sendo, pode-se, enfim, concluir que são *outliers* da amostra os perfis de número $1, 4, 15, 19, 24$ a 28. As figuras 62 e 63 trazem os diagramas de Bode com o resultado final proposto com a identificação dos perfis discrepantes para o módulo e a fase.

Tabela 16 - Resumo *outliers* e não-*outliers* para os perfis de impedância. Em negrito os perfis *outliers* que não foram detectados olhando-se separadamente o módulo e a fase.

Dados	**Não-*outliers***	***Outliers***
Módulo	$1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$ $16, 17, 18, 19, 20, 21, 22, 23, 24, 28, 29$ e 30	$4, 25, 26$ e 27
Fase	$1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16,$ $17, 18, 20, 21, 22, 23, 24, 28, 29$ e 30	$4, 15, 19, 25, 26$ e 27
Módulo e fase	$2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16$ $17, 18, 20, 21, 22, 23, 29$ e 30	$\mathbf{1}, 4, 15, 19, \mathbf{24},$ $25, 26, 27$ e $\mathbf{28}$

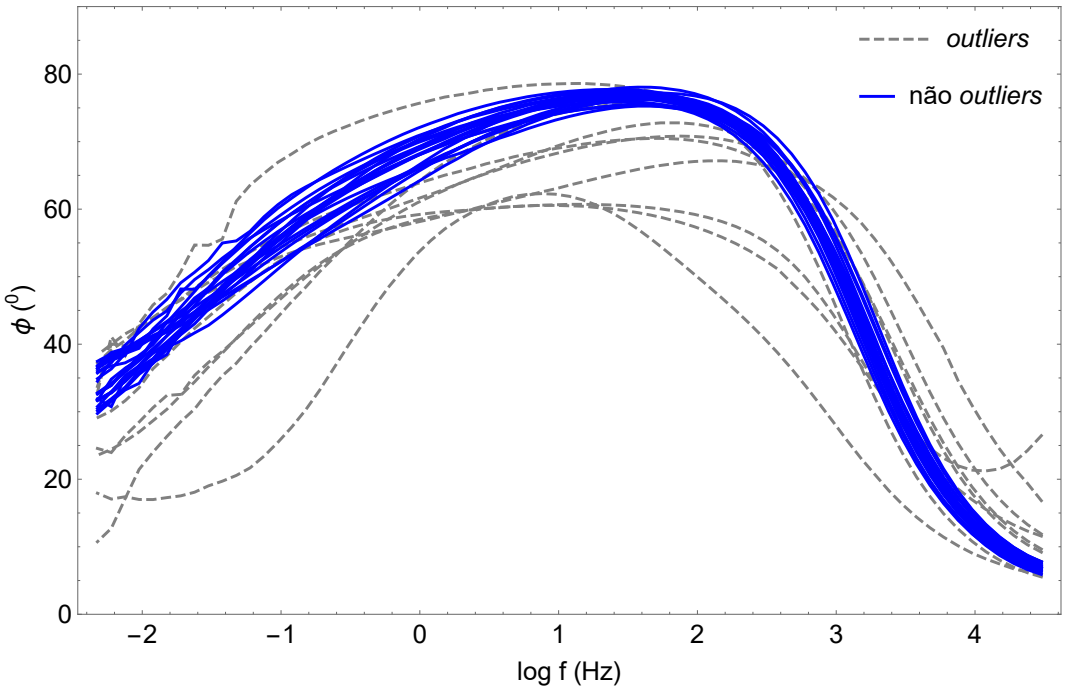
Fonte: O autor, 2022.

Figura 62 - *Outliers* e não-*outliers* no diagrama módulo *versus* log frequência.



Fonte: O autor, 2022.

Figura 63 - *Outliers* e não-*outliers* no diagrama fase *versus* log frequência.



Fonte: O autor, 2022.

5 IMPACTO DO TAMANHO DA VIZINHANÇA NOS MAPAS DE DIFUSÃO

Busca-se neste capítulo explorar ainda mais o uso do parâmetro de escala no núcleo de difusão na técnica de mapas de difusão. Por meio de testes em perfis simulados não-lineares, mostra-se que a análise do gráfico de distâncias globais a cada variação desse parâmetro exibe regiões de sensibilidade que fornece diversas informações. Isso foi visto, anteriormente, com o estudo do caso envolvendo os perfis de ensaios de corrosão. Tal gráfico, expressão visual da função denotada por M, é comparado com outros procedimentos para a escolha desse parâmetro e os resultados são promissores.

Segundo Singer e Wu (2011), um dos principais objetivos na análise de um grande conjunto de dados de alta dimensão é identificar sua estrutura geométrica e topológica. Pelo menos daqueles conjuntos que dispõe de tal estrutura, tal propriedade deve ser investigada. Neste sentido, várias técnicas exploram a redução de dimensionalidade com eficácia que busca a representação de um conjunto de dados de maior dimensão em outro menor, preservando principalmente as características essenciais. De acordo com Coifman e Lafon (2006), a nova representação deve descrever os dados de maneira fiel, preservando, por exemplo, algumas quantidades de interesse, como distâncias mútuas locais.

No decorrer do processo para atingir a redução, como no caso dos mapas de difusão, o principal passo é a definição do parâmetro de escala que definirá a noção de proximidade entre cada ponto de dados. De modo geral, esse é um desafio para diversas técnicas. Tal parâmetro remete à escala considerada em algoritmos que usam a distância euclidiana como métrica e, de outro modo, representa o limite onde se imagina que essa distância seja próxima da geodésica em estruturas não-lineares. Na clássica técnica de mapas de difusão aqui abordada, a noção de proximidade estabelecida entre pontos de dados obtidos é decisiva para a disposição final dos dados com o mapeamento. Isso foi visto, por exemplo, no caso em estudo no capítulo 3.

Se por um lado o parâmetro de escala desempenha um papel importante na técnica, por outro ele não recebe a atenção devida na literatura, sendo escolhido de forma $ad-hoc$. Como observado em alguns trabalhos que utilizam o algoritmo de mapas de difusão, a justificativa para o uso do valor especificado desse parâmetro não é clara e, se algum critério para escolha é utilizado, a maioria dos trabalhos não o apresentam. Podem-se citar como exemplos: Barkan et al. (2013), Salhov et al. (2015), Luo et al. (2015) ou Chen et al. (2016). Este capítulo procura mostrar a importância do parâmetro, usando para tal dados simulados.

5.1 **Perfis simulados**

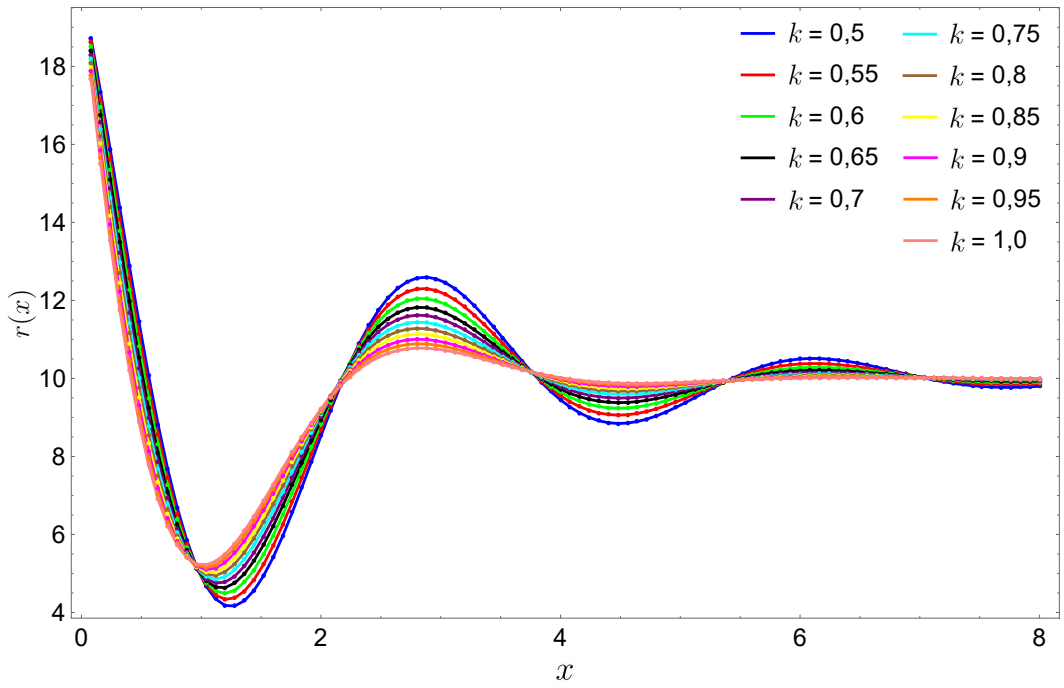
A função geratriz que fornece os perfis simulados foi utilizada como exemplo por Zhang e Albin (2009), na busca de *outliers* pelo método de cartas de controle χ^2 e, por Moura Neto, Souza e Magalhães (2019), para ilustrar a eficiência assintótica do método de mapas de difusão para fornecer uma estimativa imparcial do perfil padrão de uma linha de base. O objetivo desta função é, então, simular um conjunto de perfis não-lineares que teriam sido obtidos de algum experimento. A expectativa é que a análise em um conjunto de dados sob o qual se detêm mais controle forneça informações precisas sobre a importância do parâmetro de escala α na técnica de mapas de difusão.

Seja a família de funções,

$$r(x) = 10 - 20ke^{-kx} \cdot \frac{\sin(\sqrt{4-k^2}x)}{\sqrt{4-k^2}} + 10e^{-kx} \cdot \cos(\sqrt{4-k^2}x) \tag{37}$$

parametrizada por k. A figura 64 mostra $r(x)$ para diversos valores de k. Cada amostra de perfil simulado é um vetor das quais as entradas são números reais, utilizando $x_l = 0,08l$, com $l \in \{1, 2, 3, \ldots, 100\}$, adicionado de um ruído com desvio padrão normal $\gamma \sim N(0, \sigma^2)$.

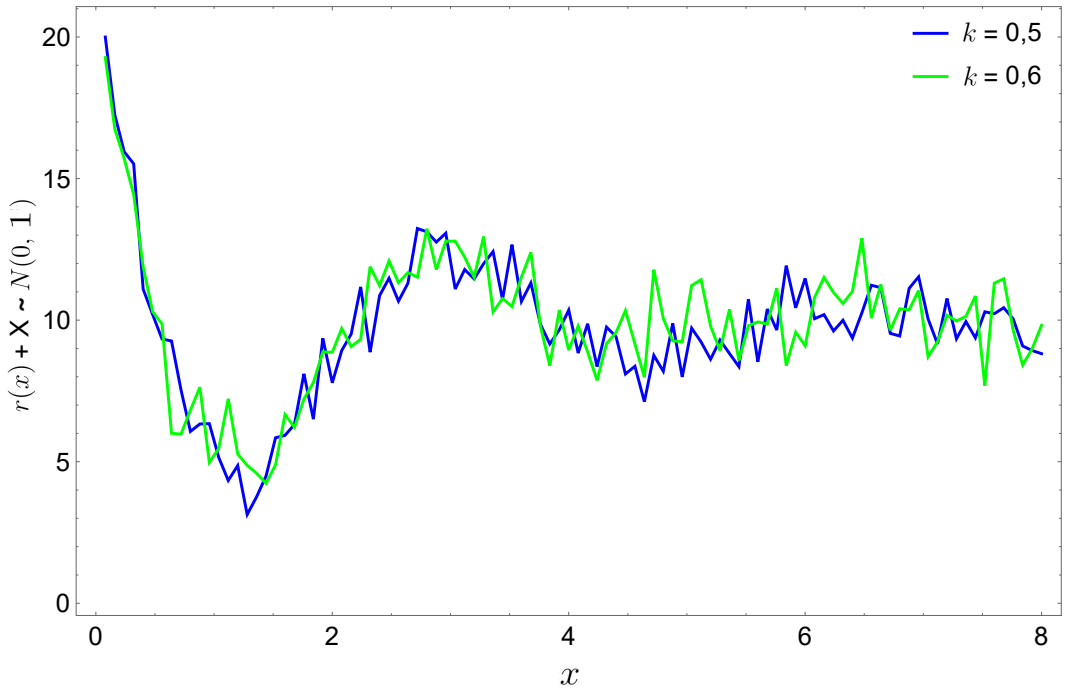
Figura 64 - Perfis base simulados com $k = 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9; 0,95$ e $1,0$.



Fonte: O autor, 2022.

A figura 65 mostra duas amostras de perfis já compostas com sua parte determinística ($k = 0,5$ e $k = 0,55$) e estocástica ($\sigma = 1$).

Figura 65 - Exemplos de perfis gerado aleatoriamente com $k = 0,5$ e $k = 0,6$.

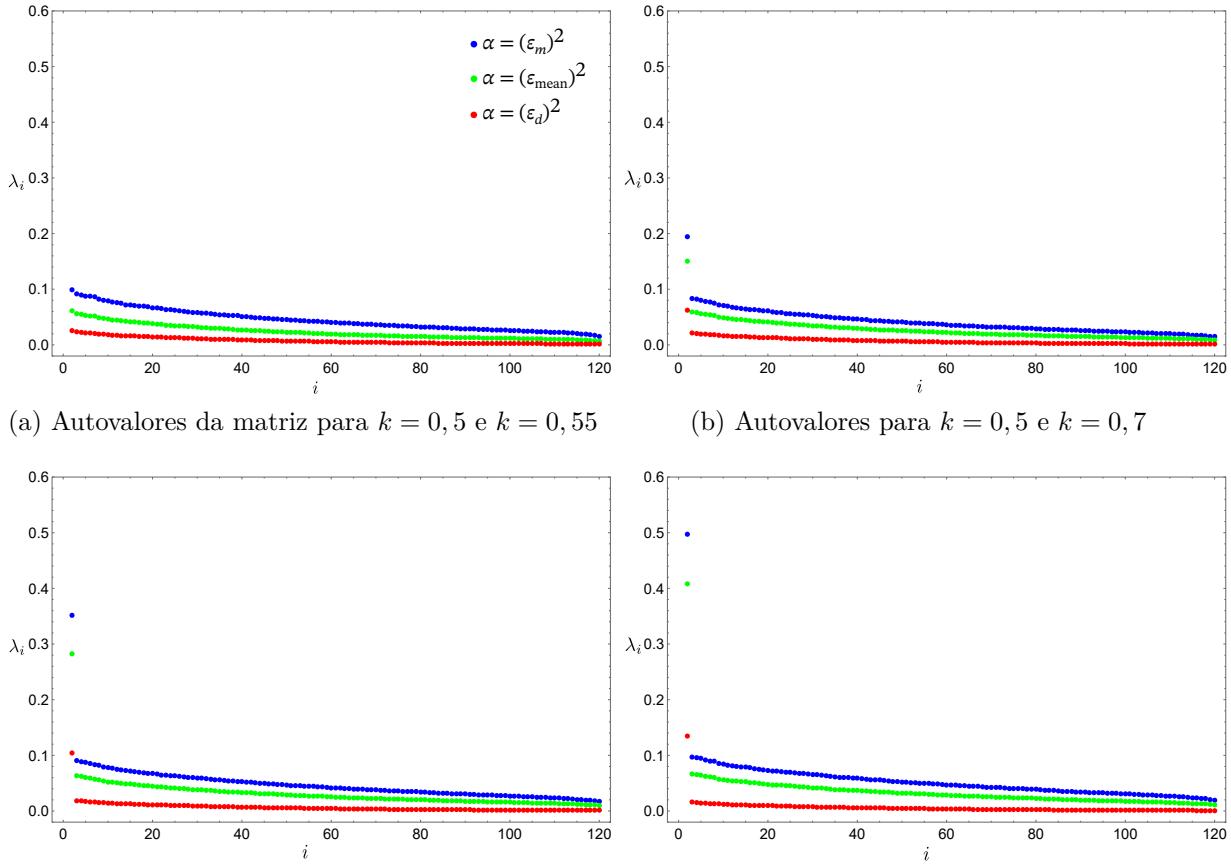


Fonte: O autor, 2022.

Para entender o efeito do parâmetro no mapeamento com os mapas de difusão, vários testes foram realizados. Em todos os resultados aqui exibidos foram considerados dois grupos de perfis simulados, cada um com 60 amostras. Fixou-se o parâmetro $k = 0,5$ para um dos grupos e variou-se esse parâmetro em passos de $0,05$ para outro grupo até alcançar $k = 1,0$. A justificativa para tal é observar como a técnica e a variação do parâmetro de escala se comportam em grupos de perfis representando *clusters* intrincados (quando os dois grupos têm valores próximos de k) e para k mais distantes.

A primeira análise com a aplicação dos mapas de difusão é o efeito da variação do parâmetro α nos autovalores da matriz de difusão. Tal escolha tem interferência significativa dentro do mesmo grupo (mantendo k fixo) e acentua-se à medida que os *clusters* são mais bem definidos (variando k). Isso será visto adiante. A figura 66 traz os autovalores da matriz de difusão P, a partir do segundo, para três situações diferentes: $\alpha = (\varepsilon_{\mathrm{m}})^2$, $\alpha = (\varepsilon_{\mathrm{d}})^2$ e um valor intermediário $\alpha = (\varepsilon_{\mathrm{mean}})^2$, com o segundo grupo assumindo $k = 0,55$; $k = 0,7$; $k = 0,85$ e $k = 1,0$.

Figura 66 - Autovalores da matriz de difusão P para diferentes grupos de perfis simulados fixando $k = 0,5$ para o primeiro grupo e variando k para o segundo. Da esquerda para direita, de cima para baixo, k assume os valores $k = 0,55$; $k = 0,7$; $k = 0,85$ e $k = 1,0$ para o segundo grupo. A cor azul representa $\alpha = (\varepsilon_{\mathrm{m}})^2$, o verde $\alpha = (\varepsilon_{\mathrm{mean}})^2$ e o vermelho, $\alpha = (\varepsilon_{\mathrm{d}})^2$.



(a) Autovalores da matriz para $k = 0,5$ e $k = 0,55$

(b) Autovalores para $k = 0,5$ e $k = 0,7$

(c) Autovalores para $k = 0,5$ e $k = 0,85$

(d) Autovalores para $k = 0,5$ e $k = 1,0$

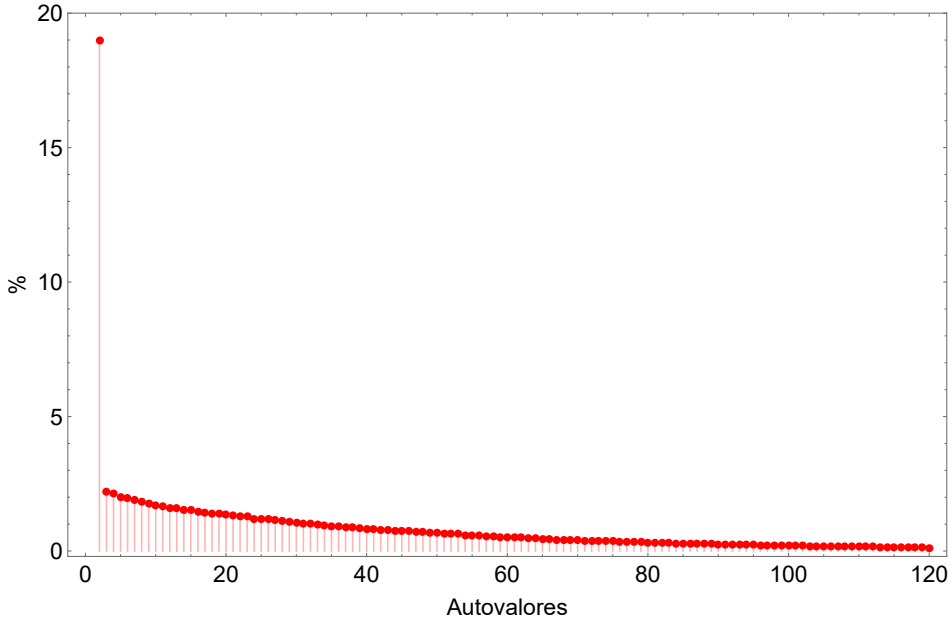
Fonte: O autor, 2022.

Analisando a figura 66a, vê-se os valores obtidos pelos autovalores da matriz de difusão com α em momentos distintos para os dois grupos mais próximos. É observado que os autovalores decaem lentamente, com o segundo autovalor (primeiro dado em cada curva) bem próximo aos demais. A interpretação extraída daqui é que todos os autovalores são importantes. Se procedermos com a redução, será difícil escolher um grupo de autovalores dominantes que serão responsáveis pelas coordenadas de cada dado mapeado. De acordo com a equação 17, cada autovalor, a partir do segundo, é responsável por uma coordenada principal no mapeamento.

Observando agora as figuras 66b, c e d, é possível notar que o valor atribuído ao segundo autovalor se destaca cada vez mais à medida que os grupos se distanciam. É como se os mapas de difusão já identificassem, por meio da matriz de difusão, que tratam-se de grupos distintos e que somente um parâmetro livre (o valor de k) diferencia os grupos.

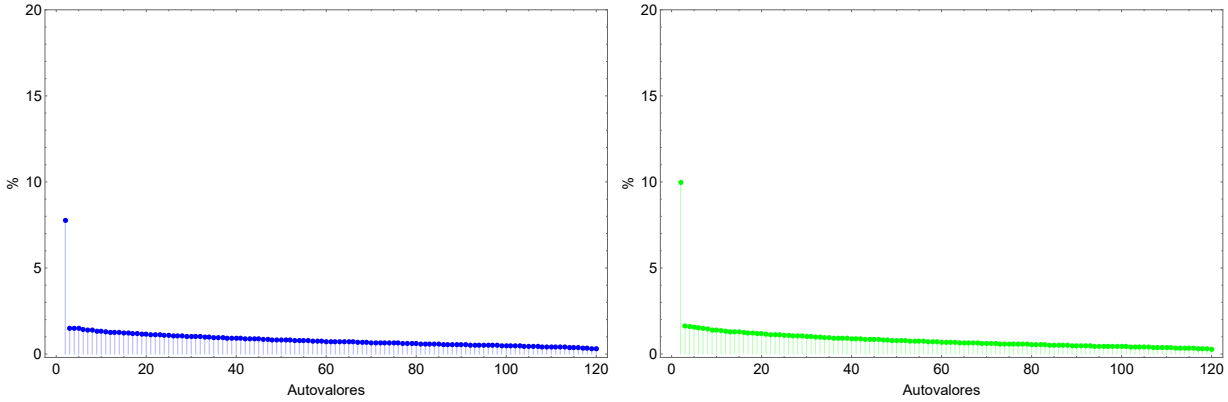
Na figura 66d, o segundo autovalor tem um peso maior em relação aos demais e chega a representar quase 20% do total (Fig. 67). Valores percentuais mostram quanto cada autovalor equivale em relação à soma na figura 68.

Figura 67 - Percentual dos autovalores para os dois grupos mais distantes ($k = 0,5$ e $k = 1,0$) com $\alpha = (\varepsilon_{\mathrm{d}})^2$.



Fonte: O autor, 2022.

Figura 68 - Percentual dos autovalores para os dois grupos mais distantes ($k = 0,5$ e $k = 1,0$) para o limite de conectividade e o parâmetro de escala médio.



(a) $\alpha = (\varepsilon_{\mathrm{m}})^2$ (b) $\alpha = (\varepsilon_{\mathrm{mean}})^2$

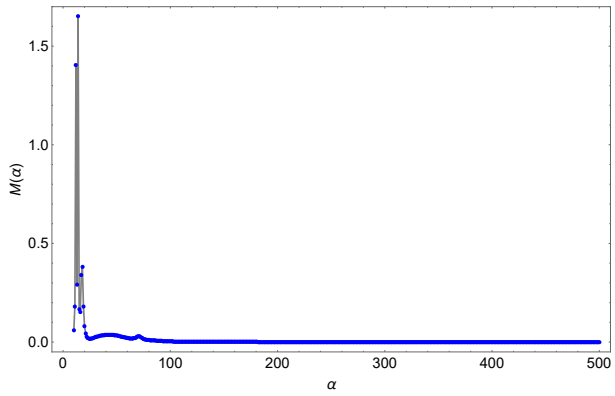
Fonte: O autor, 2022.

A análise seguinte será realizada considerando a função M para os dados simulados como definido anteriormente. Nessa abordagem, os limites do intervalo onde α é escolhido

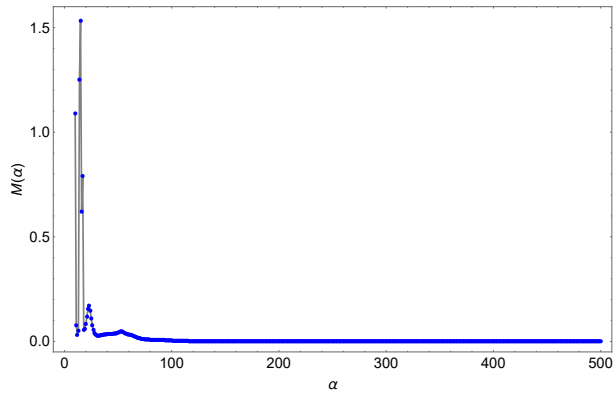
foram $a = 10$ e $b = 500$. Recorde que estes são os limite do intervalo no qual α varia. É importante destacar que, para todos os conjuntos testados (diferentes valores de k para o segundo grupo), estão incluídos o quadrado do limite de conectividade ($\alpha = (\varepsilon_{\mathrm{m}})^2$) e o quadrado do diâmetro do conjunto ($\alpha = (\varepsilon_{\mathrm{d}})^2$). Dessa forma, a intenção é que o intervalo considerado compreenda toda a região de interesse.

A escolha do número de dimensões para a redução nesta abordagem foi $m = 2$. Apesar de talvez não ser ideal para o caso onde os grupos estão intricados, o objetivo desta redução é a possibilidade de poder observar por meio de uma visualização dinâmica a relação entre o comportamento das curvas com o gráfico de $M(\alpha)$ e a disposição dos dados mapeados a cada variação. Como visto no capítulo 3, a visualização dinâmica aqui se refere a uma sequência de *frames* (vídeo) onde cada um exibe o mapeamento 2D para cada valor do parâmetro α escolhido como se esse fosse o tempo no vídeo. O passo p de variação de α foi definido em 1, assim como o parâmetro temporal t. A figura 69 traz $M(\alpha)$ para os diferentes dados abordados com k variando de $0,55$ a $1,0$ no segundo grupo.

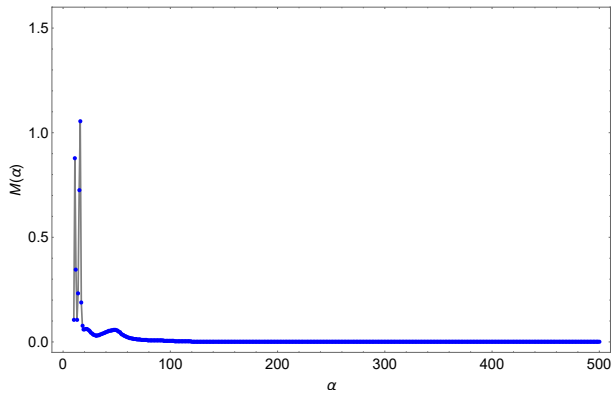
Figura 69 - $M(\alpha)$ para diferentes grupos de perfis simulados com $\alpha \in [10, 500]$.



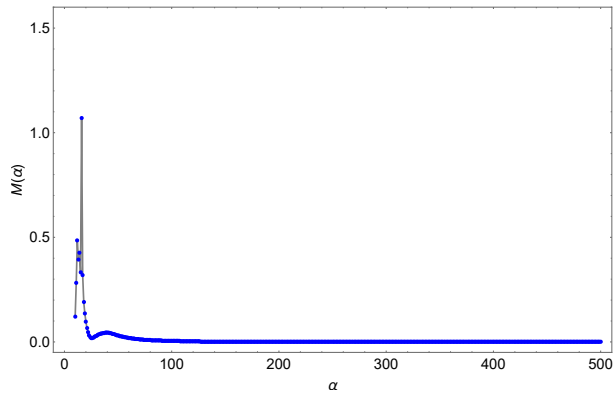
(a) $k = 0,5$ e $k = 0,55$



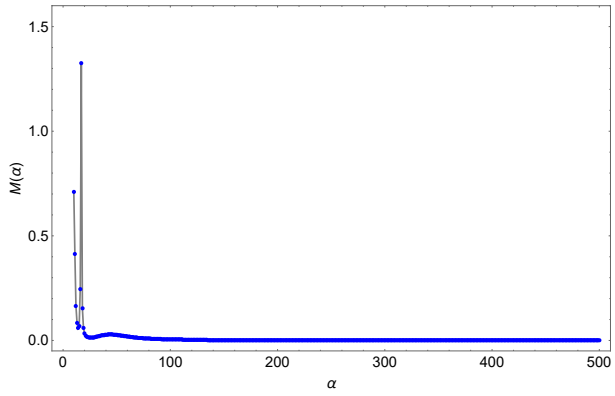
(b) $k = 0,5$ e $k = 0,6$



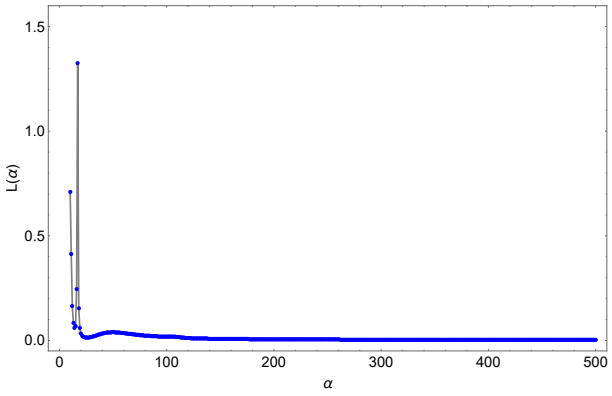
(c) $k = 0,5$ e $k = 0,65$



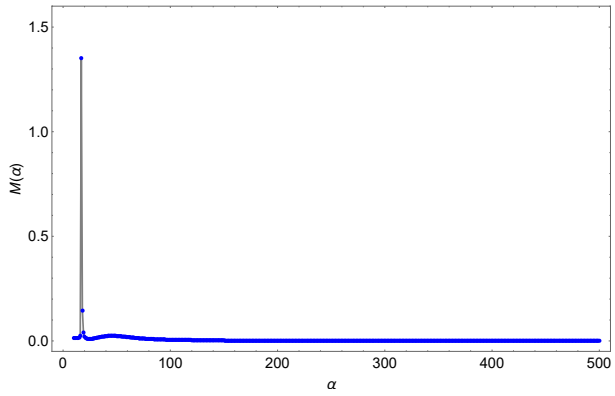
(d) $k = 0,5$ e $k = 0,7$



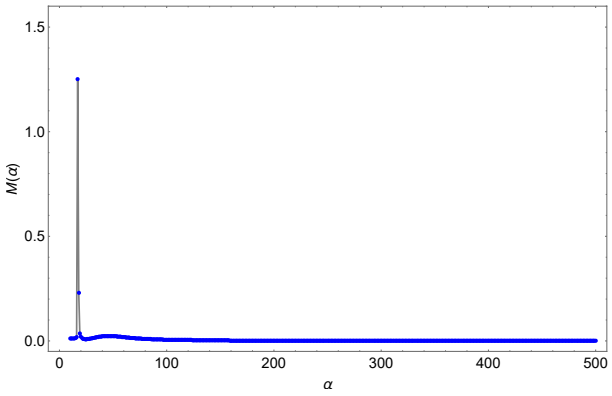
(e) $k = 0.5$ e $k = 0.75$



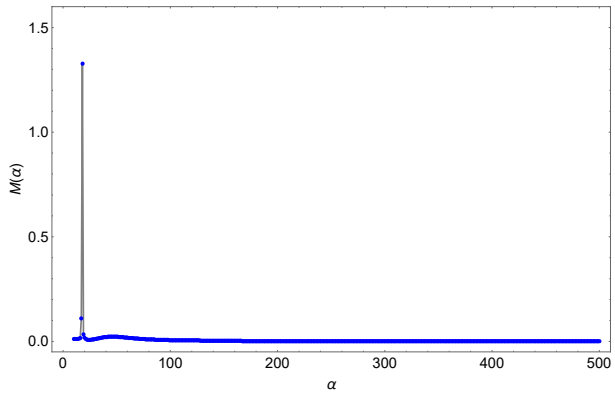
(f) $k = 0.5$ e $k = 0.8$



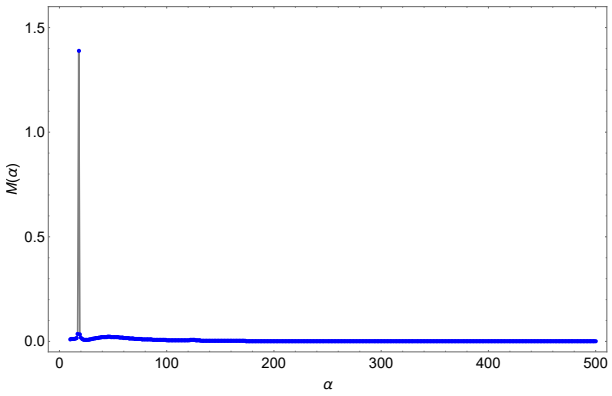
(g) $k = 0.5$ e $k = 0.85$



(h) $k = 0.5$ e $k = 0.9$



(i) $k = 0.5$ e $k = 0.95$



(j) $k = 0.5$ e $k = 1.0$

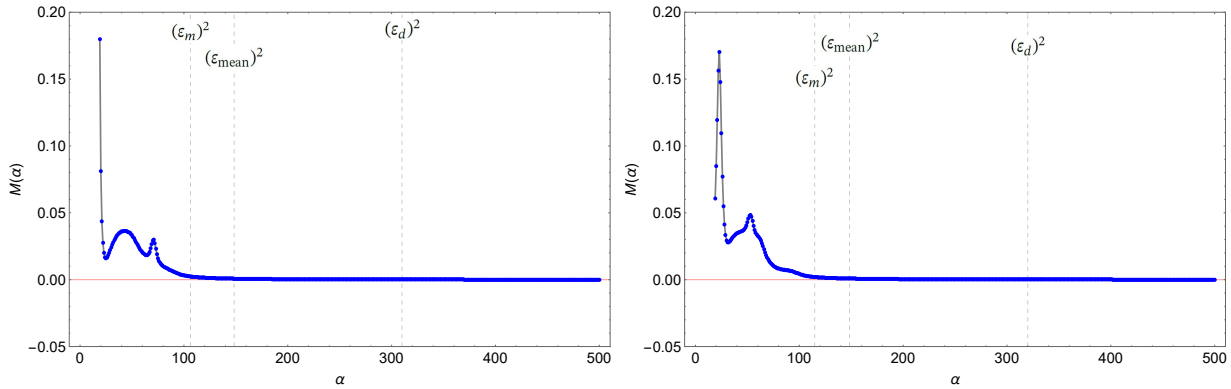
Fonte: O autor, 2022.

Fazendo a análise da figura 69, a primeira observação em destaque é a presença de uma região desordenada entre $\alpha = 10$ e $\alpha = 20$. Nessa faixa, cada mapeamento ocorre de forma quase aleatória, tendo para cada valor dentro desse intervalo, o conjunto de dados mapeados para distintos lugares do plano. A visualização dinâmica destaca esse comportamento que pode ser atribuído, talvez, ao uso de um pequeno valor para o parâmetro de escala abaixo de um certo limite suportado. Isso é visto para todos os conjuntos de dados simulados.

A explicação para este comportamento pode ser conferida à dificuldade do algoritmo em lidar com números muito pequenos. Se α tende a zero, a expressão na equação 10 tende também a zero, uma vez que $(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\alpha)$ tende ao infinito. Estes valores muito pequenos podem resultar na perda de precisão na rotina computacional com o ambiente utilizado. Cabe ressaltar que o valor de α entre 10 e 20 é inferior ao quadrado do limite de conectividade que fica próximo de $\alpha = 100$ em todos os conjuntos.

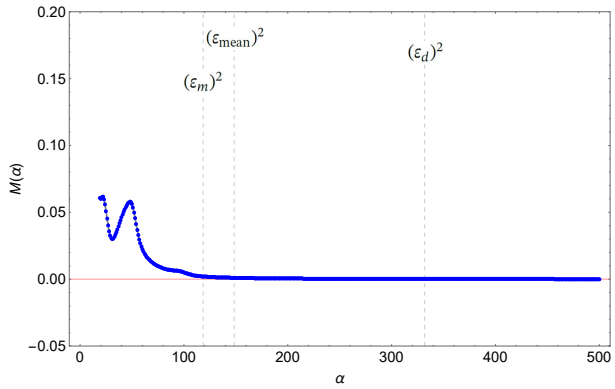
A fim de melhorar a análise para regiões acima de $\alpha = 20$, a figura 70 exibe $M(\alpha)$ com $\alpha \in [20, 500]$. Desta forma, retira-se do intervalo de estudo a região desordenada e a inspeção pode ater-se à real mudança na disposição dos dados em função do parâmetro. Como informação adicional, foram incluídas referências para cada grupo em relação aos valores do limite de conectividade, valor médio e diâmetro do conjunto.

Figura 70 - $M(\alpha)$ para diferentes grupos de perfis simulados com $\alpha \in [20, 500]$. As linhas verticais pontilhadas mostram os valores para os referenciais utilizados com os valores do limite de conectividade, valor médio e diâmetro do conjunto.

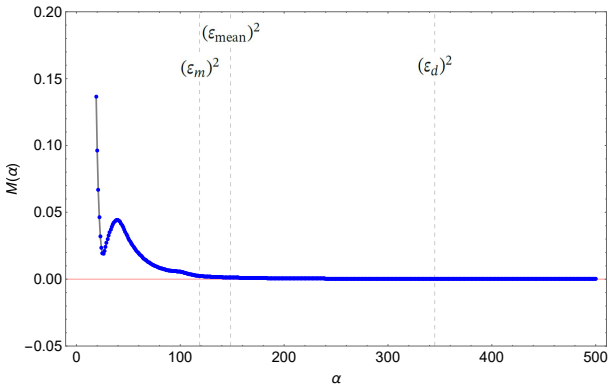


(a) $k = 0,5$ e $k = 0,55$

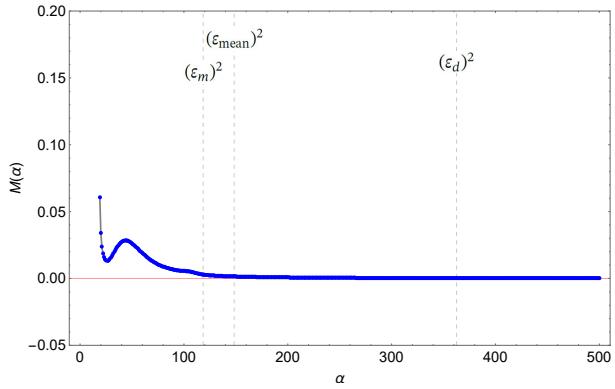
(b) $k = 0,5$ e $k = 0,6$



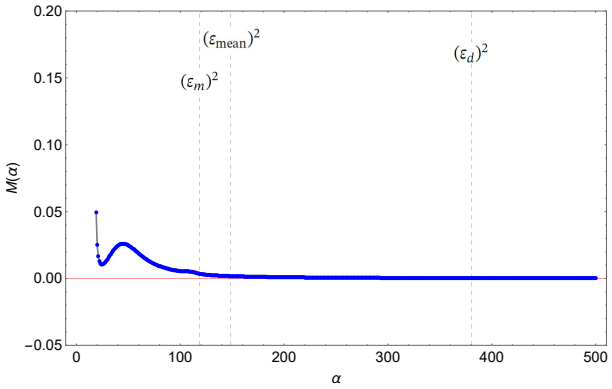
(c) $k = 0,5$ e $k = 0,65$



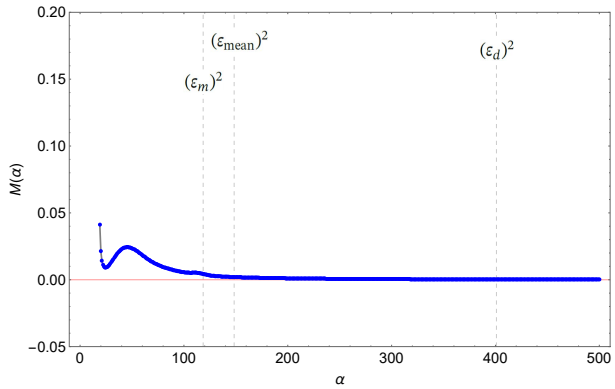
(d) $k = 0,5$ e $k = 0,7$



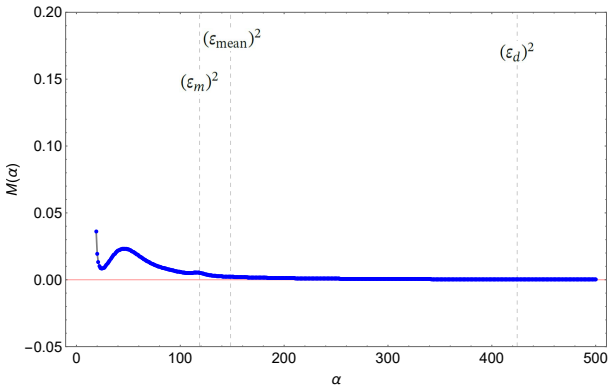
(e) $k = 0,5$ e $k = 0,75$



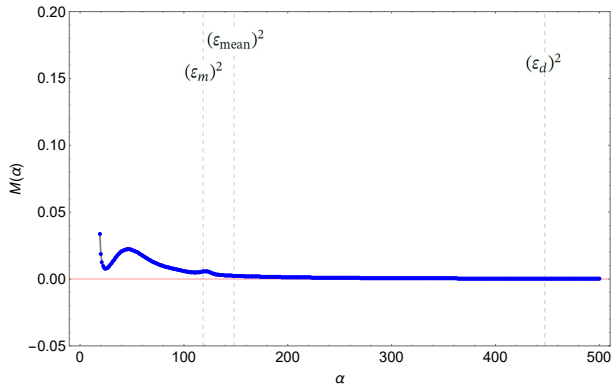
(f) $k = 0,5$ e $k = 0,8$



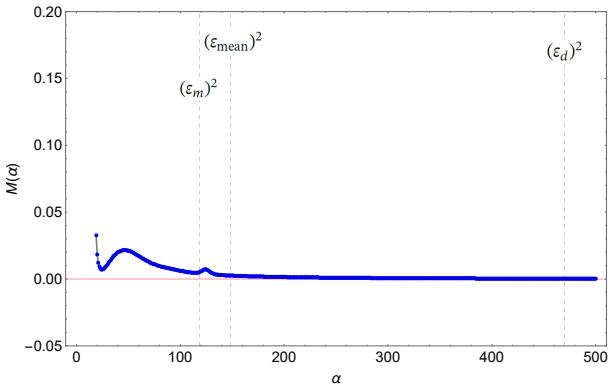
(g) $k = 0,5$ e $k = 0,85$



(h) $k = 0,5$ e $k = 0,9$



(i) $k = 0,5$ e $k = 0,95$



(j) $k = 0,5$ e $k = 1,0$

Fonte: O autor, 2022.

Fazendo agora a análise da figura 70, observa-se uma região de atividade principalmente antes de $\alpha = 120$, mais intensa quando os grupos estão intricados. Nessa faixa, a disposição dos dados mapeados é muito sensível ao parâmetro, principalmente devido à possíveis *outliers* na amostra. Ao analisar o mapeamento por meio da visualização dinâmica para esta faixa, constata-se que as maiores contribuições para a soma em $M(\alpha)$ é derivada das distâncias percorridas pelos dados mais distantes das amostra, ao passo que os pontos de dados mais similares se agrupam ao centro e se movem menos com a variação do parâmetro.

Para $\alpha \in [120, 500]$, aproximadamente, $M(\alpha)$ é decrescente e isso mostra que a disposição dos dados não mais se altera significativamente. É também possível observar que a função tende a zero à medida que α cresce. Desta forma, a alteração deste parâmetro nesta faixa não mais altera a organização dos dados mapeados e isso pode se traduzir em outra escolha não apropriada, pois temos todos os dados agora similares entre si. Se α tende ao infinito, agora, a expressão na equação 10 tende a um, uma vez que $(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / \alpha)$ tende a zero.

5.1.1 Autovalores e a formação da função $M(\alpha)$

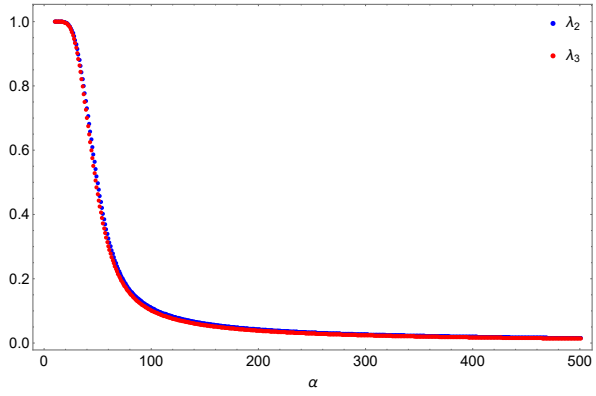
Os picos e regiões de máximo e mínimo local da função $M(\alpha)$ mostram intervalos onde o valor do parâmetro α é sensível. Os picos, como os da figura 69, podem indicar uma região de transição com um indicativo sobre o início do intervalo válido para o uso do parâmetro de escala, ou mesmo, uma região de mudança de percepção da técnica sobre a clusterização dos dados presentes. A hipótese é que a análise tem que ser feita levando em conta a vizinhança para tais valores do parâmetro que geram esse comportamento. Dessa forma, devido a ausência de pontos vizinhos que justifiquem tal ponto de máximo, os picos isolados na figura 69 têm menos importância que os relevos suaves obtidos de forma clara na figura 70, e isso justifica a análise para esse novo trecho.

Observando agora tais relevos presentes na figura 70, a suavidade destas curvas mostra se tratar de uma variação da percepção da técnica sobre o conjunto de dados diante do parâmetro de escala utilizado. Isso é confirmado por meio da visualização dinâmica que traz o mapeamento com alterações graduais e de forma sequencial, resultando na imersão dos dados. Será visto que esta região de mudança de percepção está ligada às variações dos pesos dos autovalores dominantes que geram os dados mapeados nessa faixa e isso influencia diretamente o comportamento de $M(\alpha)$.

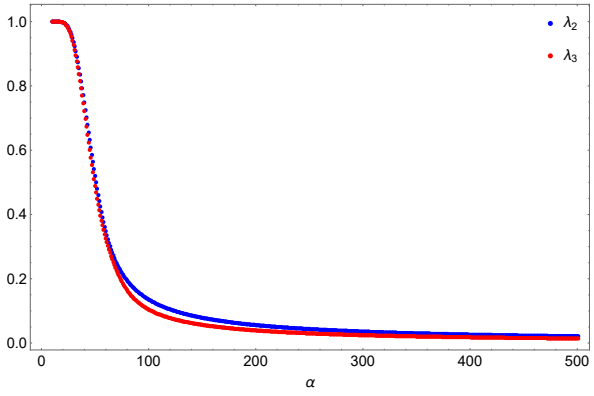
Antes disso, a próxima análise apresenta o que a figura 66 trouxe como pista. Uma vez que foi escolhido reduzir para duas dimensões, os dois maiores autovalores (com exceção do primeiro) e seus autovetores correspondentes dão as coordenadas do mapeamento. A figura 71 exibe os valores obtidos pelo segundo e terceiro autovalor dominante com a

evolução do parâmetro α para os diferentes grupos considerados. Como antes observado, o segundo autovalor ganha destaque e se sobrepõe aos demais à medida que os grupos se distanciam com o aumento de k.

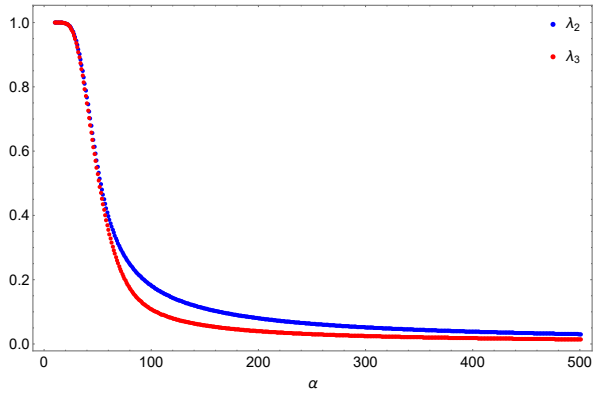
Figura 71 - Segundo e terceiro autovalores dominantes com a evolução de α para diferentes grupos de perfis simulados. O segundo autovalor se distancia do terceiro à medida que os grupos se distanciam.



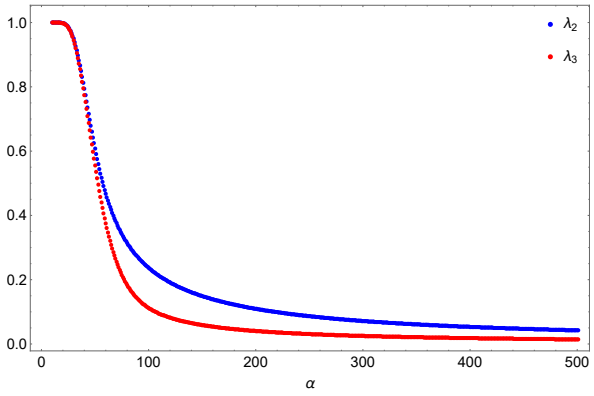
(a) $k = 0,5$ e $k = 0,55$



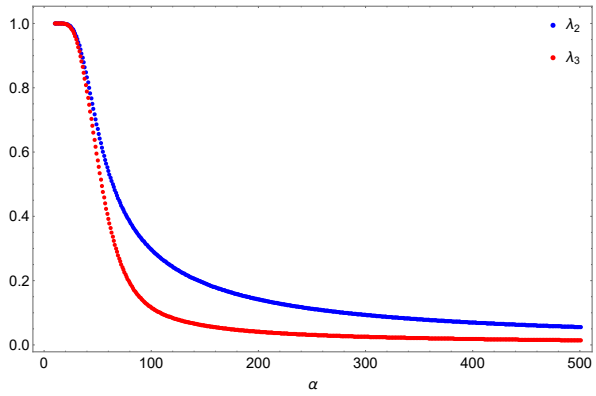
(b) $k = 0,5$ e $k = 0,6$



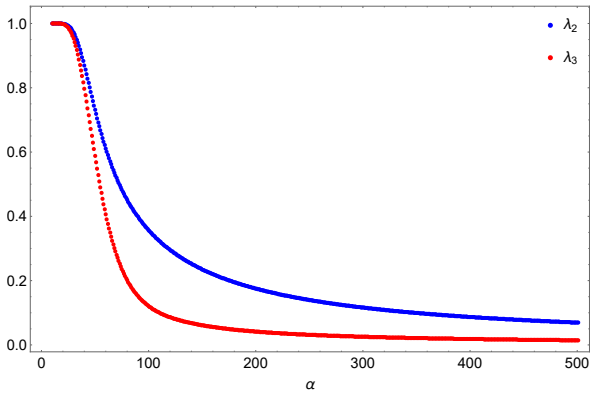
(c) $k = 0,5$ e $k = 0,65$



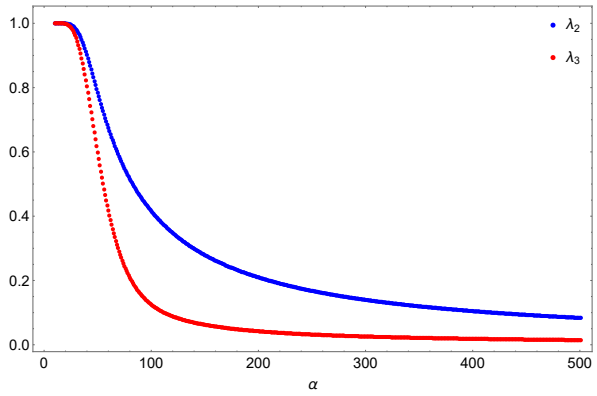
(d) $k = 0,5$ e $k = 0,7$



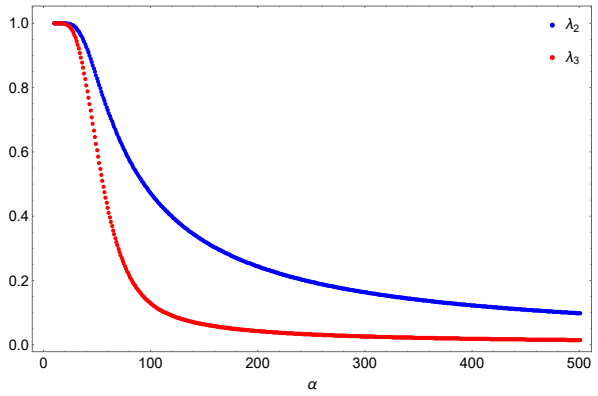
(e) $k = 0,5$ e $k = 0,75$



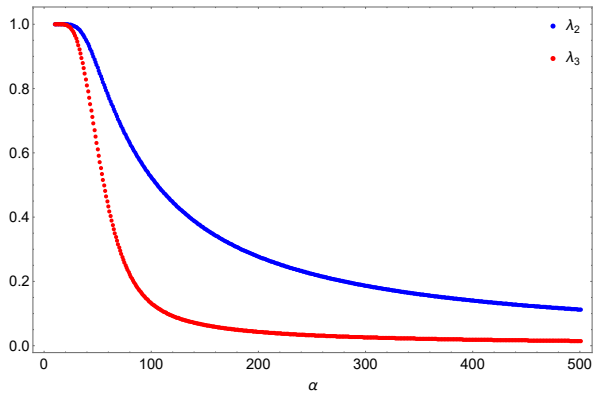
(f) $k = 0,5$ e $k = 0,8$



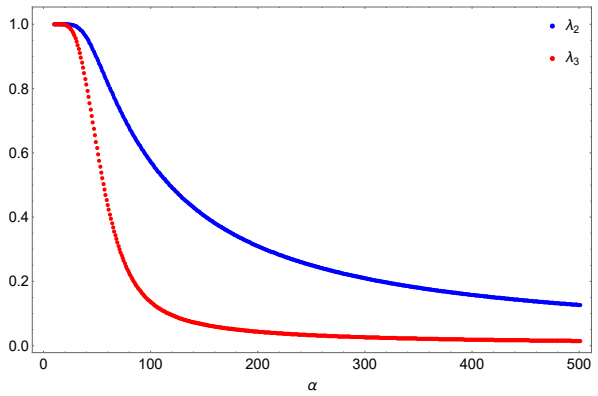
(g) $k = 0,5$ e $k = 0,85$



(h) $k = 0,5$ e $k = 0,9$



(i) $k = 0,5$ e $k = 0,95$

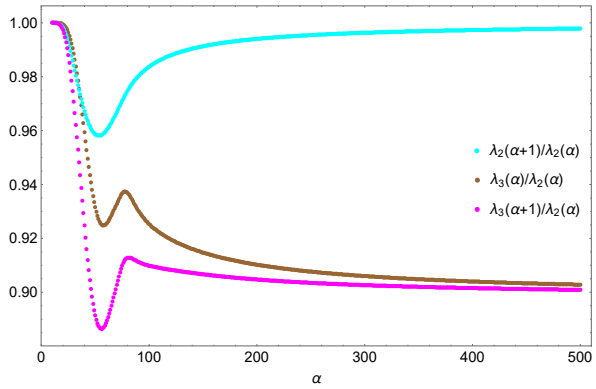


(j) $k = 0,5$ e $k = 1,0$

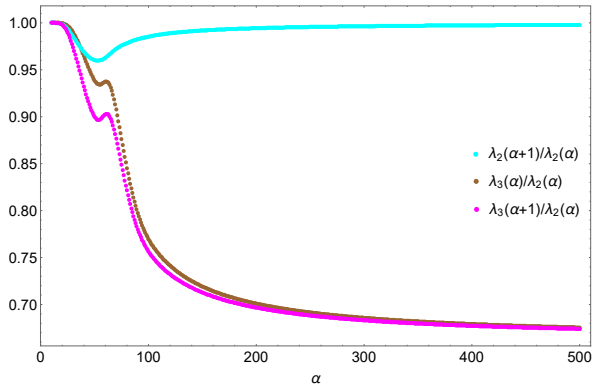
Fonte: O autor, 2022.

Observando novamente a expressão de $M(\alpha)$ na equação equação 25, verifica-se que o valor dessa função a cada alteração de α integra quatro elementos: o valor do segundo autovalor para esse parâmetro ($\lambda_2(\alpha)$), o valor do segundo autovalor para esse parâmetro em um passo adiante ($\lambda_2(\alpha+1)$), o valor do terceiro autovalor para esse parâmetro ($\lambda_3(\alpha)$) e o valor do terceiro autovalor em um passo adiante ($\lambda_3(\alpha+1)$). Como já se tem notado que o segundo autovalor é mais influente para a função M à medida que os grupos se distanciam, a figura 72 é útil para exibir que fração do segundo autovalor dominante os outros elementos constituem a cada variação de α. É visto que isso é importante para entender os relevos suaves presentes em $M(\alpha)$.

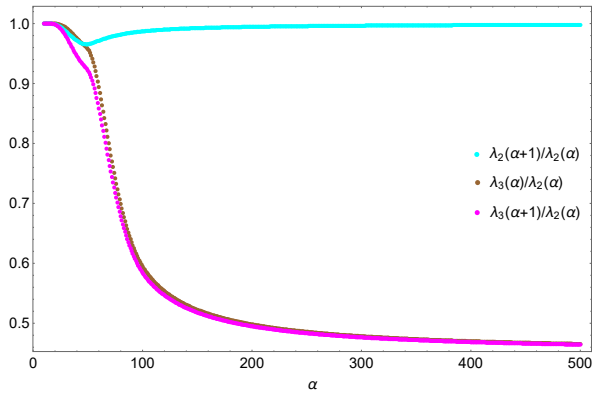
Figura 72 - Razões entre autovalores para diferentes grupos e escolhas do parâmetro α.



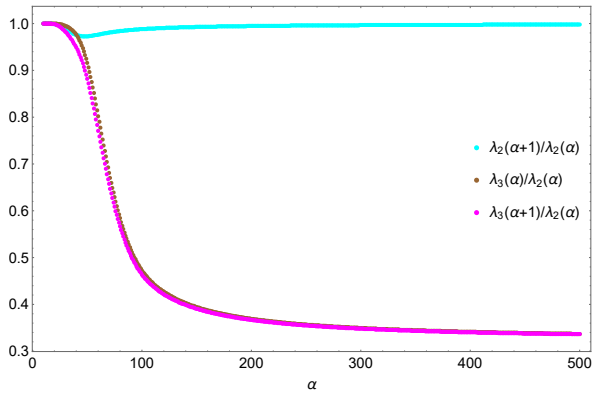
(a) $k = 0,5$ e $k = 0,55$



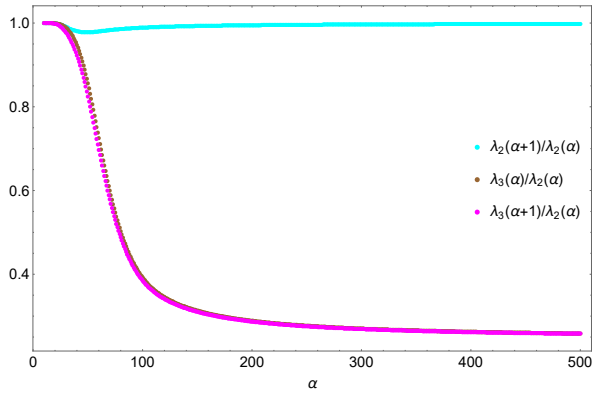
(b) $k = 0,5$ e $k = 0,6$



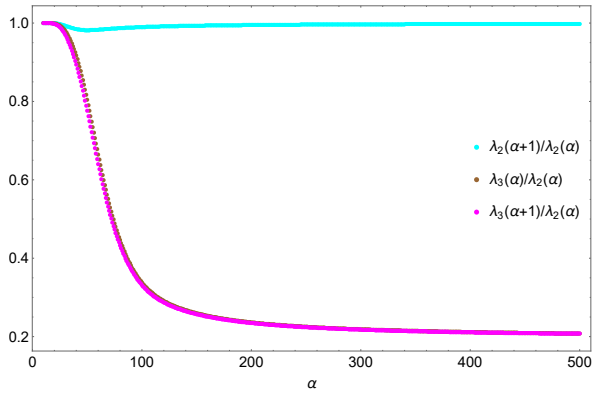
(c) $k = 0,5$ e $k = 0,65$



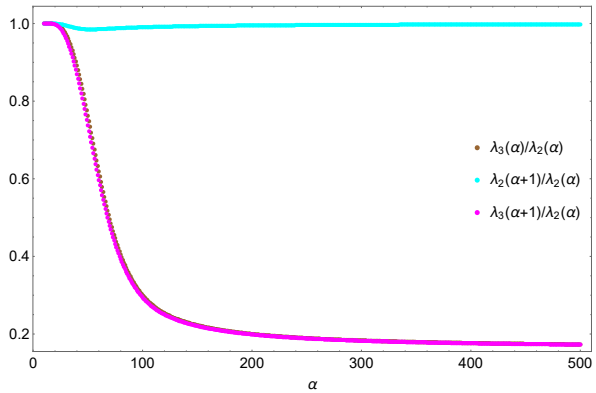
(d) $k = 0,5$ e $k = 0,7$



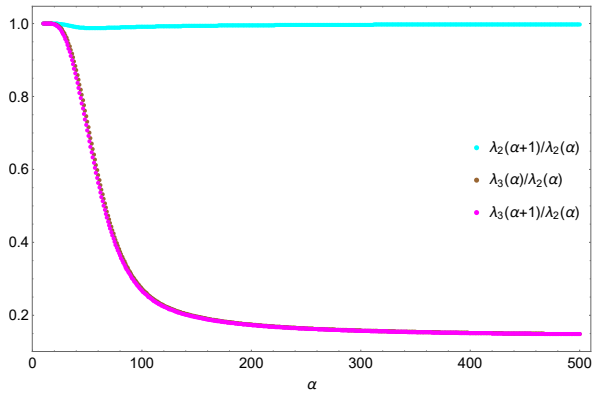
(e) $k = 0,5$ e $k = 0,75$



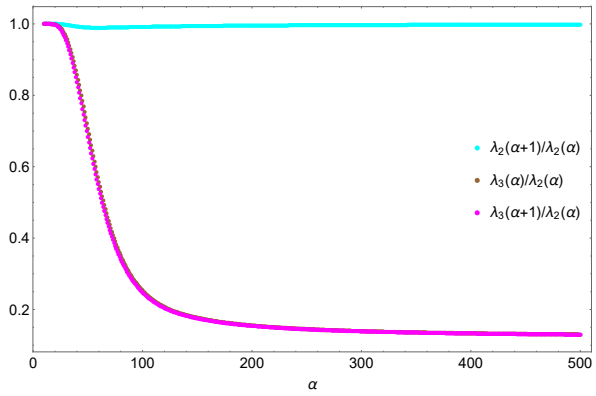
(f) $k = 0,5$ e $k = 0,8$



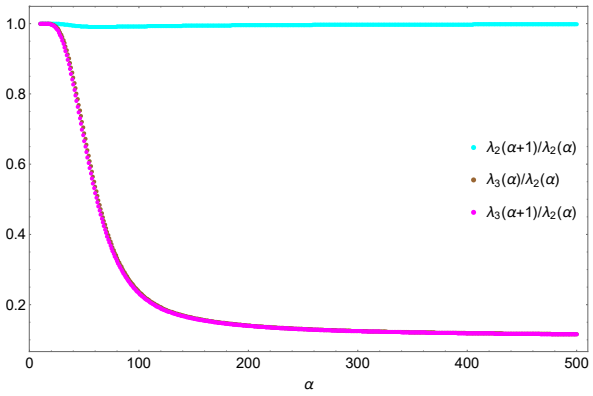
(g) $k = 0,5$ e $k = 0,85$



(h) $k = 0,5$ e $k = 0,9$



(i) $k = 0,5$ e $k = 0,95$



(j) $k = 0,5$ e $k = 1,0$

Fonte: O autor, 2022.

A análise da figura 72 trouxe ainda mais informações sobre a participação dos autovalores nas coordenadas do mapeamento. Como as subfiguras exibem, o terceiro autovalor varia de 90% na figura 72a para menos de 10% na figura 72j do que representa o segundo autovalor. Por consequência, isso evidencia que $M(\alpha)$ para os grupos mais distantes é todo motivado pelo segundo autovalor.

Outro aspecto interessante é o trecho com $20 < \alpha < 120$. Aliás, o mesmo intervalo em que a figura 70 exibia os relevos anteriormente. Observe que há uma alteração significativa na participação dos autovalores para essa faixa. É como se nesta faixa os mapas de difusão avaliassem a participação de cada uma das coordenadas importantes para o número de variáveis latentes representativas para a redução de dimensionalidade. Como argumentado anteriormente, a hipótese é que os relevos da figura 70 mostram a mudança de percepção da técnica na definição do número de autovalores necessários para representar o número de variáveis latentes do conjunto que, de outra forma, são responsáveis pela redução significativa almejada. Ao final desse trecho que, para o caso, $\alpha > 120$, tem-se uma região segura para a escolha do parâmetro. Em todos os conjuntos apresentados, a escolha de α entre o limite de conectividade e parâmetro de escala médio seria uma ótima faixa para a classificação.

5.1.2 Escolha do parâmetro α

5.1.2.1 Distribuição logarítmica de $L(\varepsilon)$

Em Singer et al. (2007), é apresentada uma maneira de escolher o parâmetro de escala do núcleo de similaridade. Na notação apresentada, chamam esse parâmetro de ε, mas é o mesmo aqui representado por α. O artigo é um dos poucos encontrados que versam sobre o tema. O trabalho de Bah (2008), por exemplo, emprega o método, como

em Singer et al. (2007), para a escolha do parâmetro. O objetivo desta seção é comparar o método proposto por meio da análise de $M(\alpha)$ com o descrito por Singer et al. (2007) para a escolha do parâmetro de escala α.

De acordo com Singer et al. (2007), o valor para o parâmetro não é único. Existe uma gama de valores possíveis dentro de uma faixa aceita para um determinado conjunto de dados. Um valor muito baixo tornaria cada dado não semelhante a qualquer outro e, do outro lado, um valor alto os tornariam todos similares. É justamente o uso da faixa de valores adequado que é buscado discutir no capítulo. Com base nessa ideia, Singer et al. (2007) propuseram o seguinte esquema:

1. Construa matrizes de similaridade dependente do parâmetro α para diferentes valores escolhidos;

2. Calcule

$$L(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij}(\alpha); \tag{38}$$

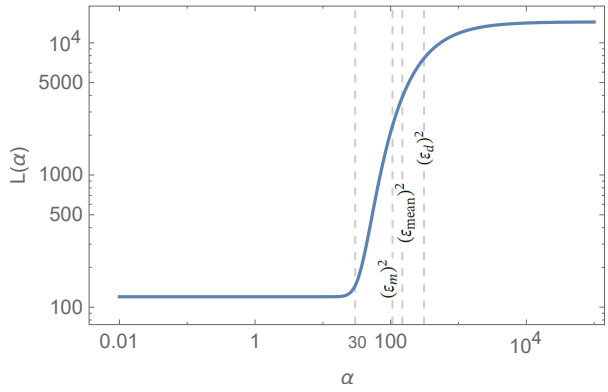
3. Obtenha a curva $L(\alpha)$ usando uma distribuição logarítmica. Esta curva terá duas assíntotas quando $\alpha \to 0$ e $\alpha \to \infty$;

4. Escolha α onde a curva $L(\alpha)$ aparece linear.

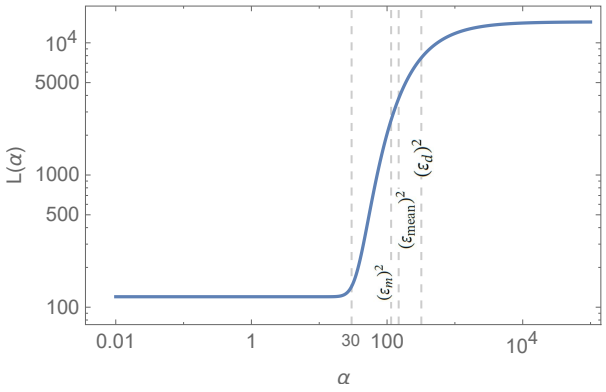
A partir deste procedimento, as curvas da figura 73 foram obtidas. Nestas acrescentou-se referências para cada grupo em relação aos valores do limite de conectividade, valor médio, diâmetro do conjunto e uma referência para o início da parte linear. Conclusões posteriores mostrarão que a análise pela função $M(\alpha)$ é mais refinada, pois aponta para uma região segura da escolha do parâmetro à medida que fornece informações adicionais sobre a estrutura dos dados envolvidos.

Segundo Bah (2008), deve-se escolher o parâmetro de escala no valor médio aproximado entre as duas assíntotas. Dessa forma, em qualquer uma das subfiguras da figura 73, a escolha do parâmetro estaria aproximadamente com $30 < \alpha < (\varepsilon_{\mathrm{d}})^2$. Singer et al. (2007), inclusive, sugerem a escolha onde a curva aparece linear. Se esta orientação é seguida, no entanto, o valor escolhido estaria com $30 < \alpha < (\varepsilon_{\mathrm{m}})^2$ que, como visto, ainda corresponde a uma região de transição de percepção da técnica caracterizada pelo relevo de $M(\alpha)$. Como consequência, o mapeamento obtido pode ser prematuro e não representar ainda a disposição dos dados de modo adequado.

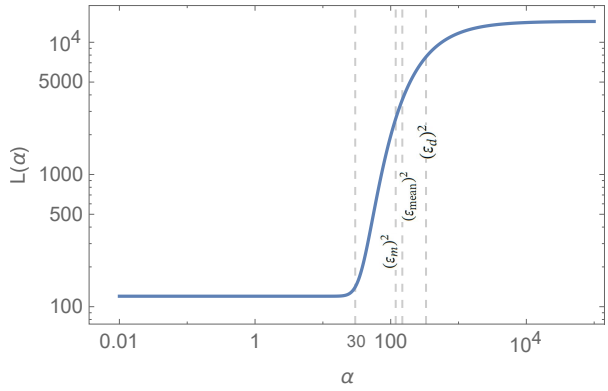
Figura 73 - Distribuição logarítmica para o conjunto de perfis simulados para diferentes valores de k.



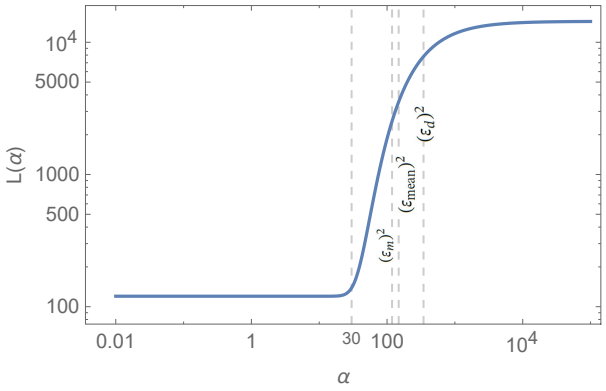
(a) $k = 0,5$ e $k = 0,55$



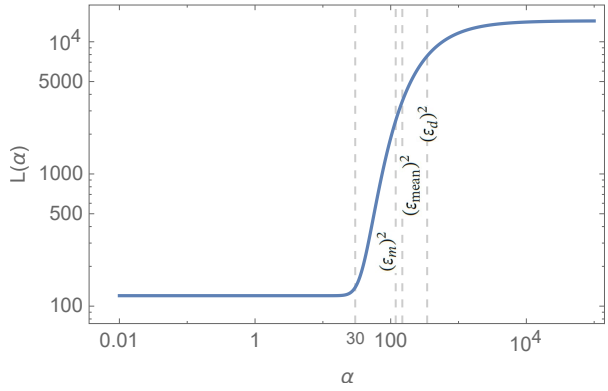
(b) $k = 0,5$ e $k = 0,6$



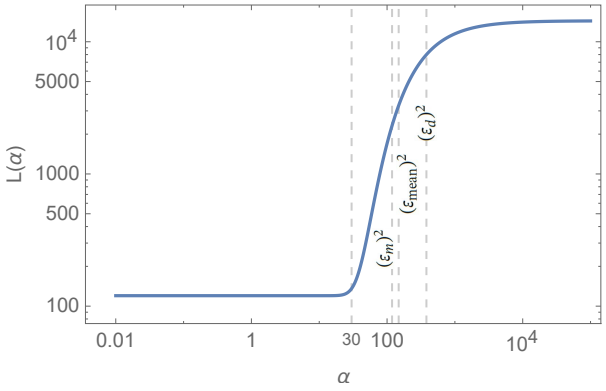
(c) $k = 0,5$ e $k = 0,65$



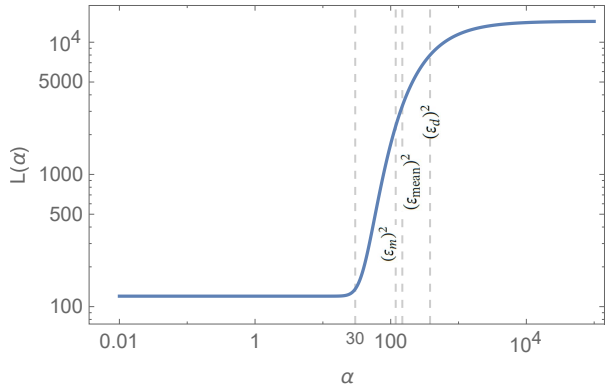
(d) $k = 0,5$ e $k = 0,7$



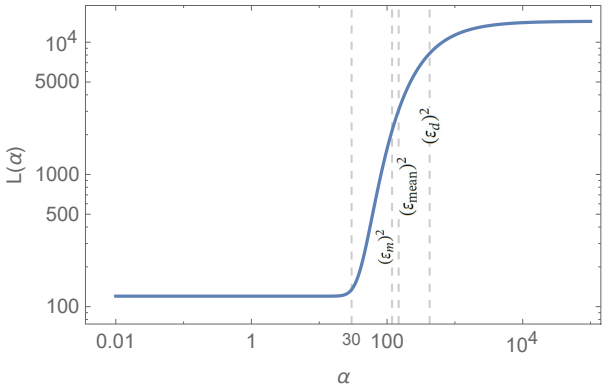
(e) $k = 0,5$ e $k = 0,75$



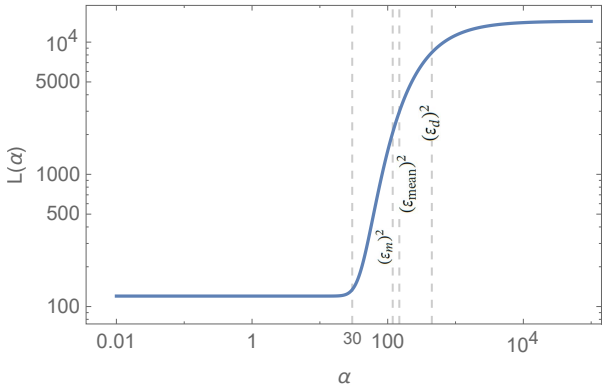
(f) $k = 0,5$ e $k = 0,8$



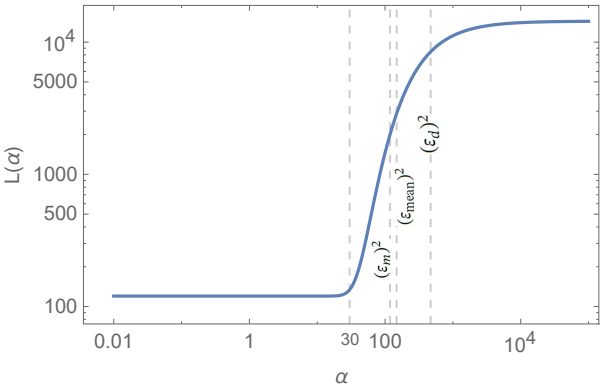
(g) $k = 0,5$ e $k = 0,85$



(h) $k = 0,5$ e $k = 0,9$



(i) $k = 0,5$ e $k = 0,95$



(j) $k = 0,5$ e $k = 1,0$

Fonte: O autor, 2022.

Outra questão sobre como o gráfico de $M(\alpha)$ é mais útil, é justamente a possibilidade de fornecer mais informações sobre o mapeamento em cada faixa de escolha do parâmetro. Se aliado à visualização dinâmica do desenrolar do mapeamento a cada mudança, permite entender como o parâmetro α afeta diferentemente cada conjunto de dados em estudo e viabilizar uma escolha mais adequada para o que se deseja. Em outro conjunto de dados, por exemplo, a mudança de percepção da técnica evidenciada pelos relevos e vales poderia estar localizada em regiões distintas. Isso não seria possível de identificar por meio do procedimento sugerido por Singer et al. (2007), uma vez que, dentro da faixa sugerida para a escolha do parâmetro, as diferentes opções podem levar a mapeamentos muito distintos.

Por fim, a análise usando a função $M(\alpha)$ mostra maiores detalhes na evolução dos mapeamentos em função de α que pode indicar diferentes regimes de análise dos dados consoante a escala (α) da descrição dos dados.

5.1.2.2 Teste do semigrupo

Em um recente trabalho publicado por Shan e Daubechies (2022) é apresentado outra maneira para a escolha do parâmetro de escala (α) do núcleo de difusão. Conforme os autores, eles aplicam a propriedade do semigrupo (*semigroup property*) do operador de difusão e propõe um critério de semigrupo (*semigroup criterion*) para a escolha "correta" do valor desse parâmetro. Nesse trabalho, esse critério é utilizado em alguns exemplos envolvendo dados sintéticos e reais, onde demonstram que diferentes escolhas do parâmetro podem levar a diferentes mapeamentos, assim como já debatido neste trabalho. O procedimento sugerido será descrito a seguir.

Segundo Shan e Daubechies (2022), a imersão eficaz dos dados residentes do espaço de recurso para o espaço de difusão pode ficar comprometida se os autovetores não forem

obtidos com precisão, principalmente, de amostras ruidosas dos dados. Por esta razão, eles julgam útil, para determinar (aproximações a) estes autovetores, trabalhar com a noção de semigrupo de operadores $\{e^{-h\Delta_{\mathscr{M}}}\}_{h\geq 0}$, com a justificativa de que este operador de difusão se relaciona com o operador Laplace-Beltrami $\Delta_{\mathscr{M}}$ por meio da equação do calor (ou difusão) em uma variedade. O h aqui é a notação utilizada para o parâmetro de escala do núcleo de similaridade pelos autores e, nesta tese, usa-se α.

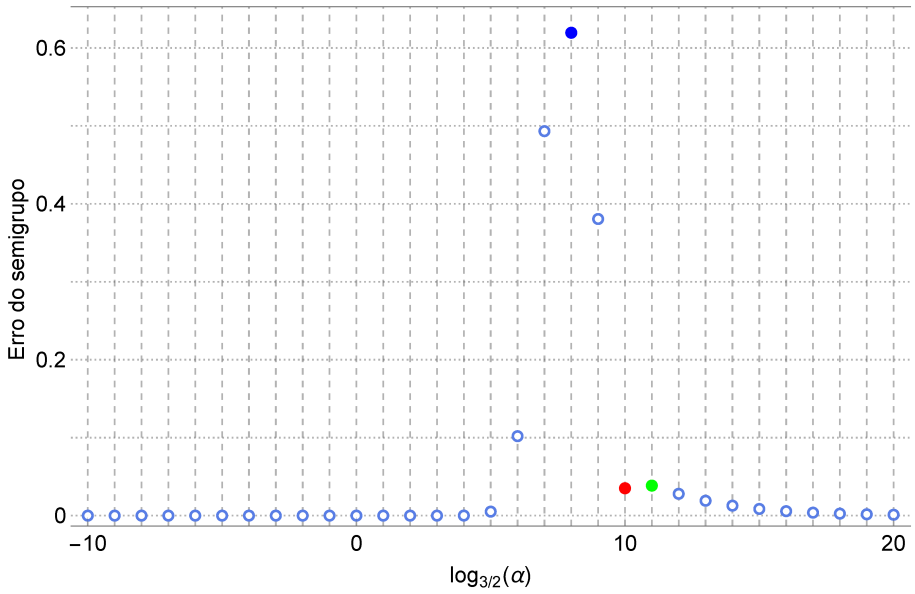
Com o argumento de que as matrizes P, a matriz de difusão, e sua versão simetrizada $K^* = Y^{-\frac{1}{2}}KY^{-\frac{1}{2}}$ podem ser versões aproximadas e discretizadas dos operadores de difusão em $\mathscr{M}$ somente quando também (aproximadamente) satisfazem à propriedade do semigrupo, eles definem o seguinte número real chamado erro do semigrupo (*semi-group error :* $\mathrm{SGE}(\alpha)$),

$$\mathrm{SGE}(\alpha) = \left\| (K^*_\alpha)^2 - K^*_{2\alpha} \right\|. \tag{39}$$

Na prática, assume-se o parâmetro de escala α em uma ampla faixa de valores, constroem-se as matrizes de difusão $(K^*_\alpha)^2$ e $K^*_{2\alpha}$ e calcula-se o $\mathrm{SGE}(\alpha)$. De acordo com o resultado proposto, o menor SGE obtido para o α utilizado dentro de uma faixa de altos e baixos valores obtidos é o valor ideal para o parâmetro. A seguir, é mostrado como o procedimento ficaria aplicado aos dados simulados em questão, abordados no capítulo.

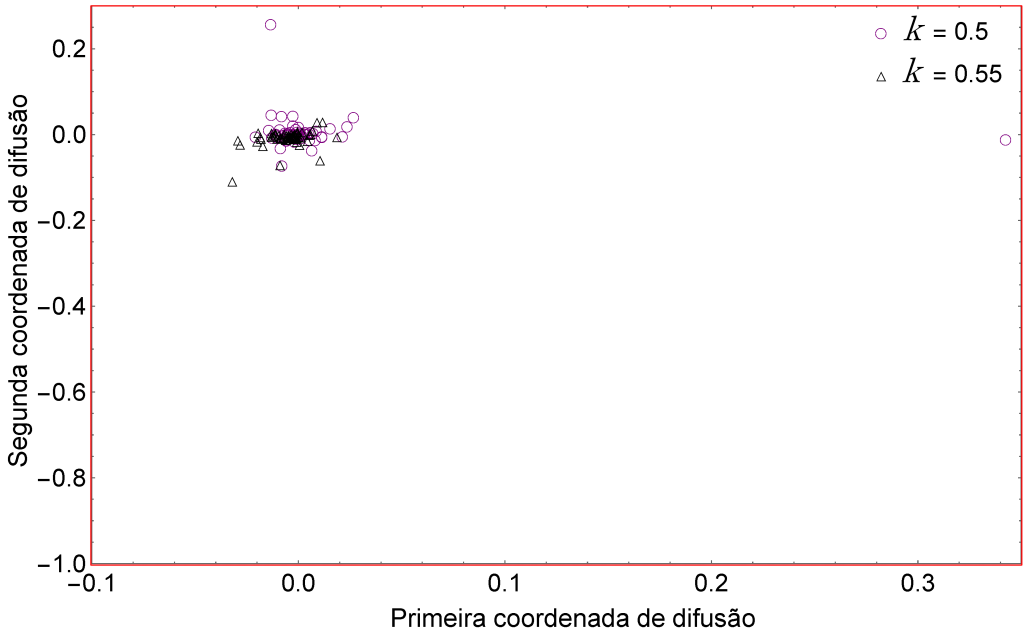
Na figura 74 é apresentado o erro de semigrupo para estes dados e, nas figuras 75 a 77, três imersões possíveis considerando alguns valores para o parâmetro α, inclusive, para o valor "correto" do parâmetro, segundo Shan e Daubechies (2022). Na figura 74, considerou-se o caso com $k = 0,55$ para o segundo grupo, simulando *clusters* menos definidos. De acordo com a metodologia, o valor apropriado para o parâmetro α é próximo de 60 ($\approx (3/2)^{10}$), em vermelho, pois equivale ao menor SGE dentro do que os autores chamam de regime de pequenos tempos de difusão.

Figura 74 - Erro do semigrupo para o conjunto de perfis simulados para $k = 0,5$ e $k = 0,55$.



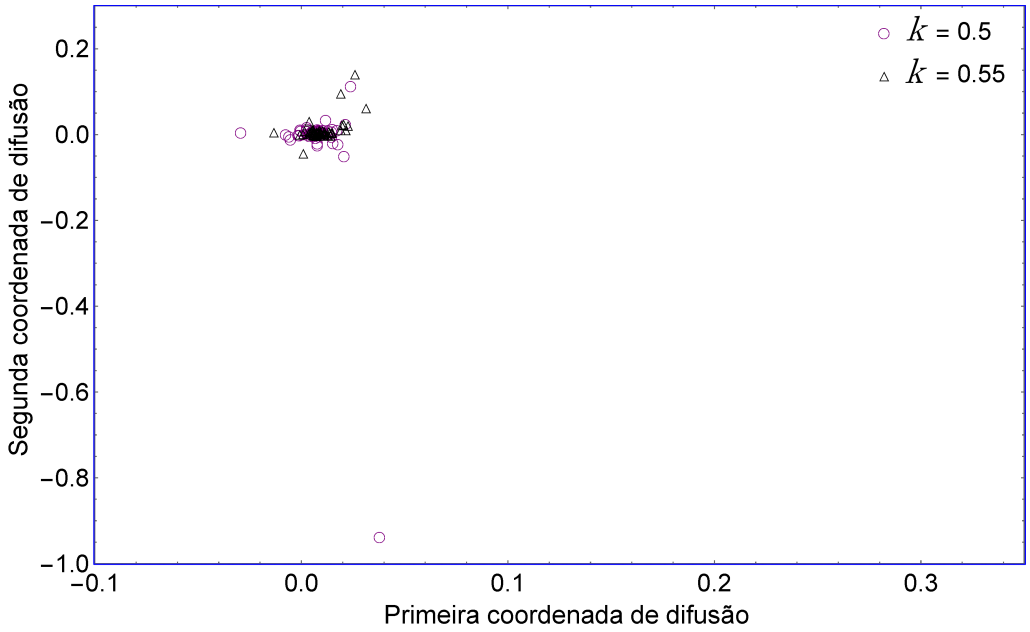
Fonte: O autor, 2022.

Figura 75 - Mapeamento 2D usando $k = 0,5$ e $k = 0,55$ para $\alpha = (3/2)^{10}$, valor sugerido pelo método de Shan e Daubechies (2022).



Fonte: O autor, 2022.

Figura 76 - Mapeamento 2D usando $k = 0,5$ e $k = 0,55$ para $\alpha = (3/2)^8$, valor ainda considerado impróprio segundo Shan e Daubechies (2022).

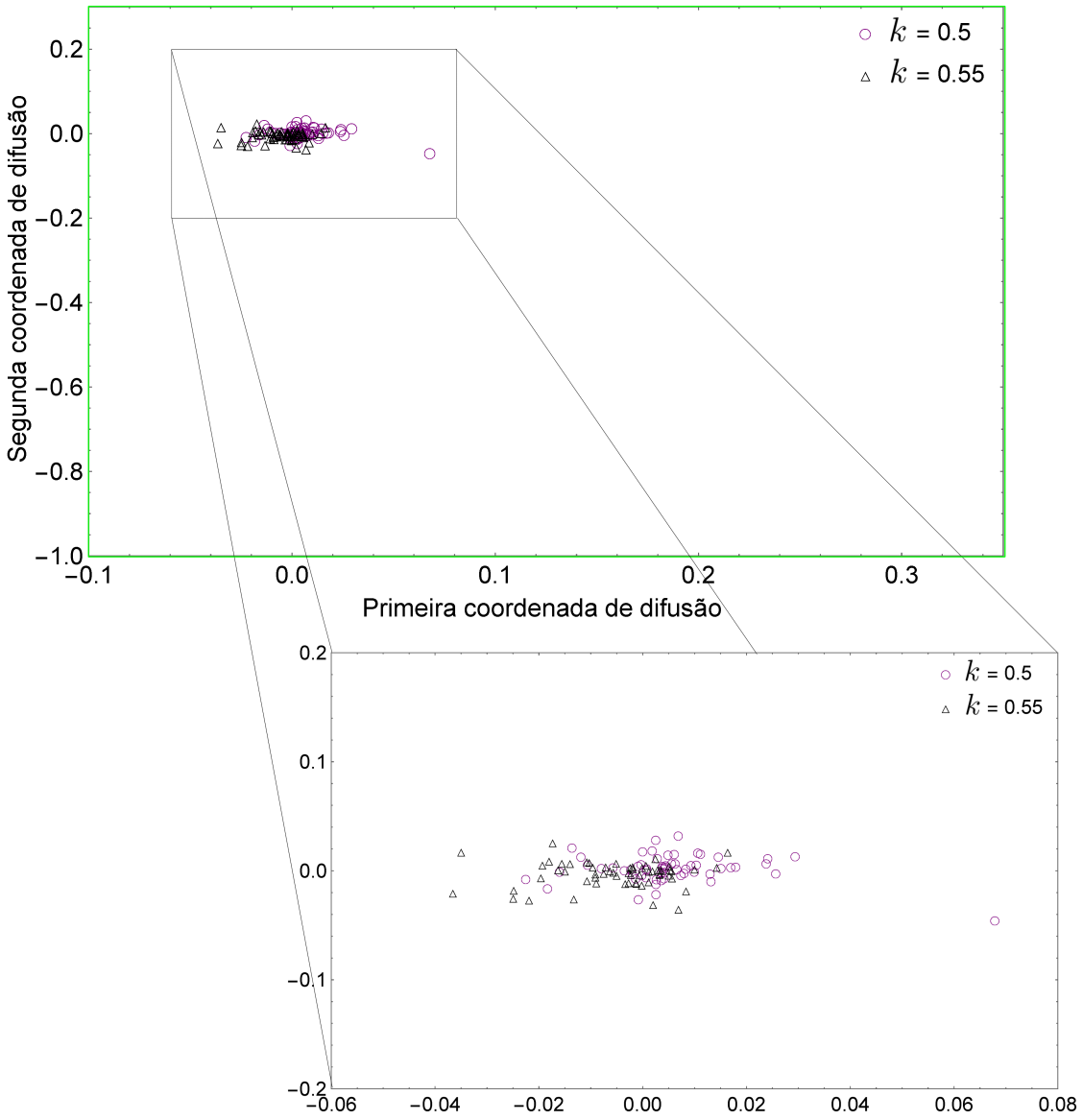


Fonte: O autor, 2022.

Ao se comparar os mapeamentos obtidos nas figuras 75 a 77, observa-se que os dois grupos de perfis simulados parecem estar melhores determinados na figura 77, apesar dos *clusters* ainda bem intrincados. Aliás, dada a proximidade dos valores para o parâmetro k para os grupos, não é possível esperar algo diferente com o mapeamento. Nas figuras 75 e 76, entretanto, os grupos parecem menos definidos, ainda, com *outliers* mapeados mais distante com $k = 0,5$. O mapeamento na figura 77 mostra que esses *outliers* foram absorvidos para o grupo, com um deles apenas mais afastado.

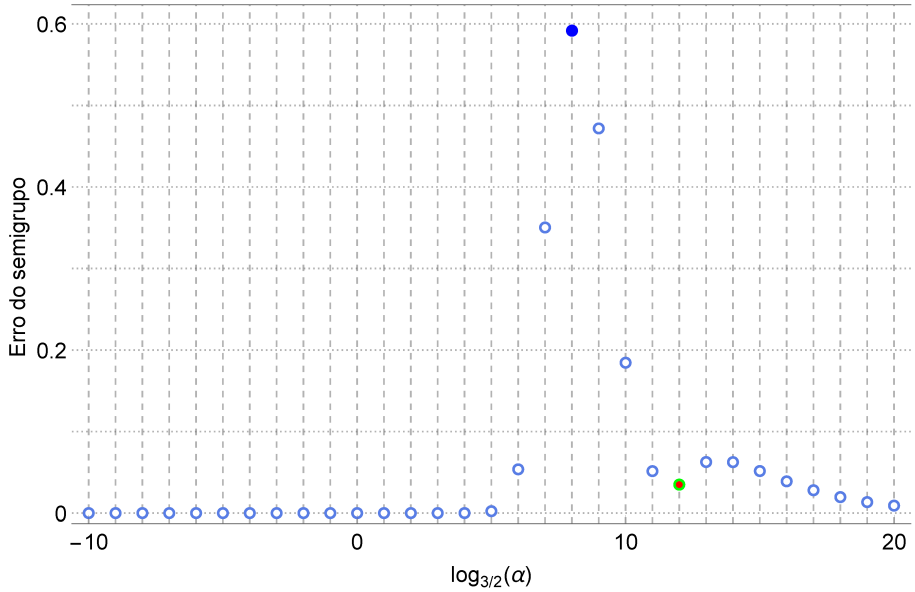
A figura 78, por sua vez, exibe o de erro de semigrupo para os mesmos dados considerando o caso com $k = 1,0$ para o segundo grupo, simulando *clusters* mais definidos. As figuras 79 e 80 trazem duas imersões possíveis considerando alguns valores para o parâmetro α, novamente, incluindo o valor sugerido por Shan e Daubechies (2022). Neste enfoque, o valor apropriado para o parâmetro α seria próximo de 130 ($\approx (3/2)^{12}$), em vermelho, seguindo o procedimento já descrito anteriormente. O mapeamento para este valor sugerido é exibido na figura 80. Para tal abordagem, entretanto, este valor do parâmetro α sugerido pelos autores coincidiu com o valor proposto neste trabalho com a análise de $M(\alpha)$, como pode ser observado na figura 70j, confirmando, mais uma vez, a utilidade do procedimento proposto para a escolha deste parâmetro de modelagem.

Figura 77 - Mapeamento 2D usando $k = 0,5$ e $k = 0,55$ para $\alpha = (3/2)^{11} \approx 87$, valor sugerido por este trabalho para os dados simulados em questão. Os *outliers* foram absorvidos para o grupo com $k = 0,5$, com um deles apenas mais distante.



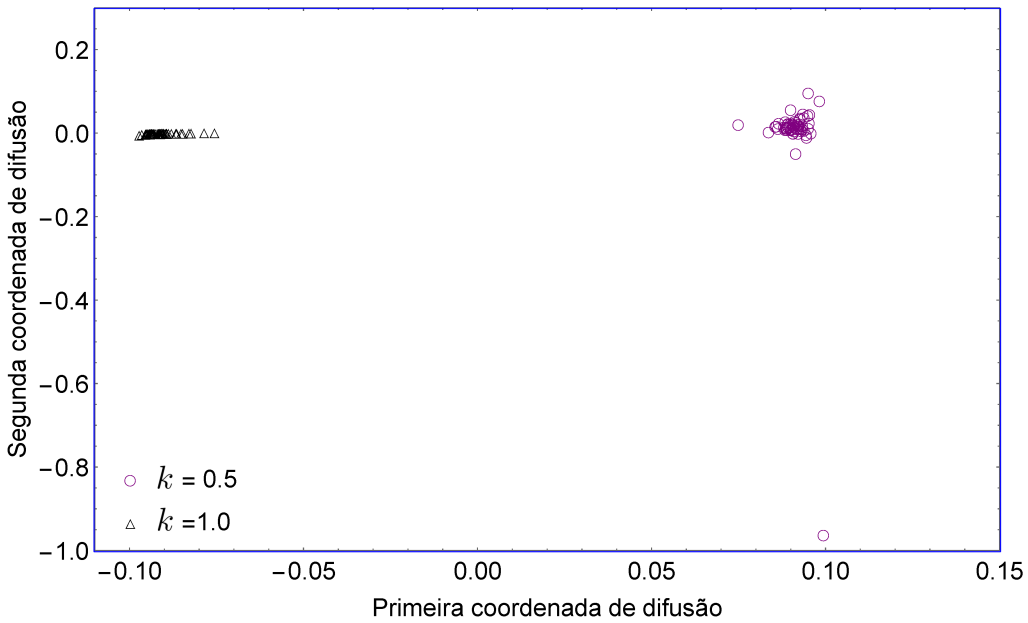
Fonte: O autor, 2022.

Figura 78 - Erro do semigrupo para o conjunto de perfis simulados para $k = 0,5$ e $k = 1,0$.



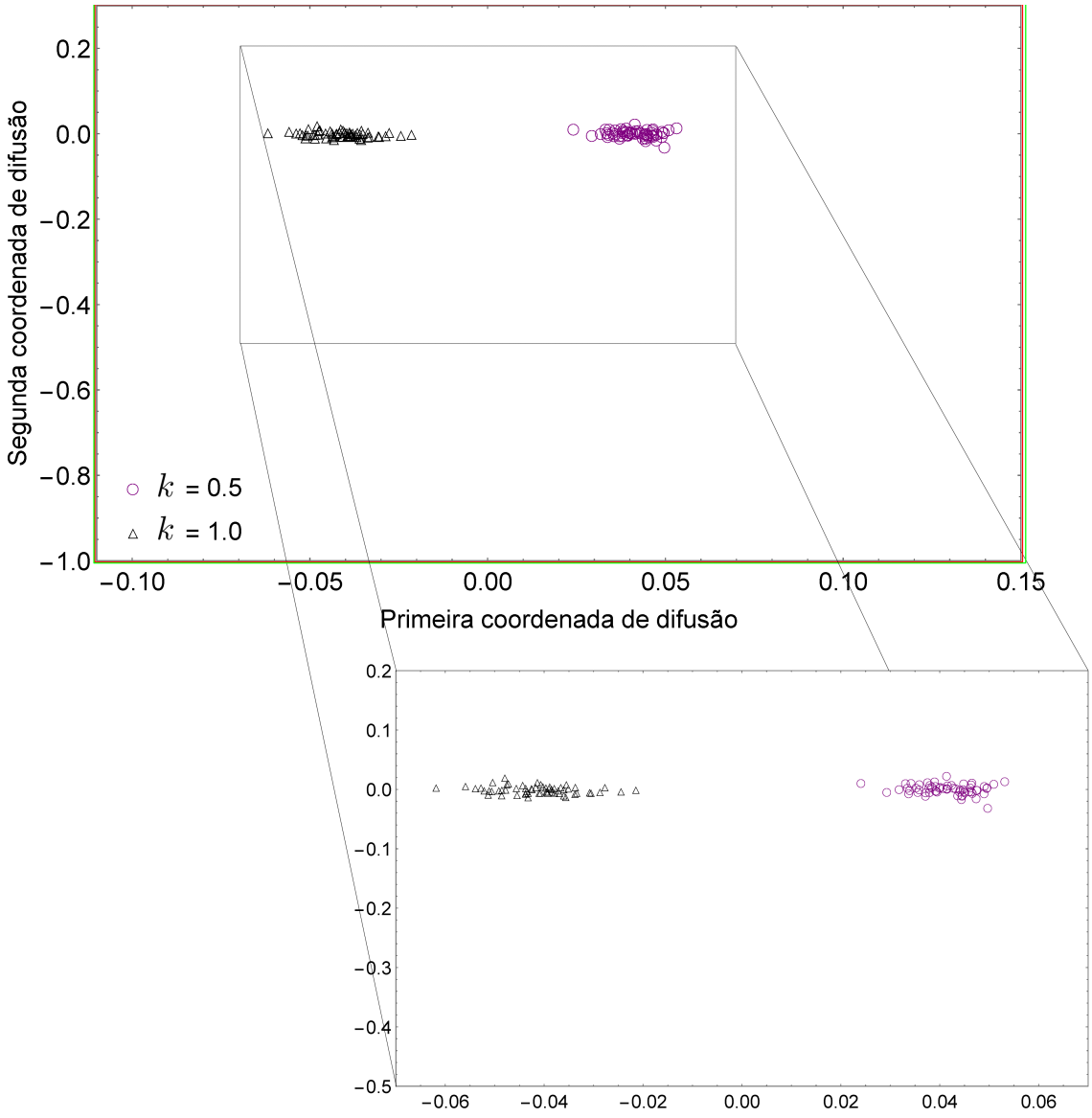
Fonte: O autor, 2022.

Figura 79 - Mapeamento 2D usando $k = 0,5$ e $k = 1,0$ para $\alpha = (3/2)^8$.



Fonte: O autor, 2022.

Figura 80 - Mapeamento 2D usando $k = 0,5$ e $k = 1,0$ para $\alpha = (3/2)^{10}$, valor sugerido pelo método de Shan e Daubechies (2022) e também proposto pelos resultados desta tese.



Fonte: O autor, 2022.

CONCLUSÕES

Nesta tese, o emprego de mapas de difusão foi empregado para classificar sinais de natureza eletroquímica como curvas de polarização potenciodinâmica e dados de espectroscopia de impedância eletroquímica. Em relação às curvas de polarização, os aços inoxidáveis austenítico UNS S30400 e S31600 em meio aquoso contendo $3,5\%$ (massa) NaCl a $25°C$ foram estudados. Estes aços diferenciam entre si pelo teor de molibdênio, que aumenta a resistência à corrosão localizada do S31600. Estes dados se apresentam como perfis não-lineares, e foram produzidas 48 réplicas de cada teste para permitir a análise realizada nesta tese, em especial, no uso de mapas de difusão. Para gerar os dados de espectroscopia de impedância eletroquímica, o material usado foi o aço inoxidável superdúplex UNS S32750 com 30 réplicas. Além dos dados experimentais, também foram empregados dados simulados sem ligação com dados de natureza eletroquímica.

A técnica de mapas de difusão se mostrou eficaz no mapeamento dos perfis representantes das curvas de polarização com o classificador *Bayes*, e se revelou promissora na classificação eficaz, robusta, sistemática e automática de curvas de perfis não-lineares, provenientes de sinais eletroquímicos.

Uma aplicação introdutória de organização de imagens digitais serviu de motivação para o uso dos mapas de difusão mostrou-se capaz de organizar eficientemente um conjunto específico de imagens digitais, ainda que apresentasse ruídos (como sombras, imperfeições na medida dos ângulos), reconhecendo padrões da imagem.

Em comparação com métodos clássicos da literatura (PCA, LLE e *isomap*) para a redução de dimensionalidade e classificação eficaz do conjunto de dados, o mapa de difusão apresentou ótimo desempenho nas taxas de classificação. Assim, esta técnica foi usada na classificação de dados obtidos de curvas de polarização de aços inoxidáveis, principalmente na faixa interesse de alto potencial. A noção de similaridade entre os dados em diversas escalas com relação ao parâmetro α revelou que este tem fundamental importância no núcleo de difusão do mapeamento do conjunto de perfis. Entretanto, a escolha adequada do parâmetro foi necessária. O núcleo gaussiano foi usado nesta análise sendo α e t os parâmetros que o ajusta.

É possível afirmar que a técnica de mapas de difusão foi útil na identificação de *outliers* dos dados experimentais expressos ao mostrar possíveis perfis discrepantes por meio da matriz de difusão. Nesta abordagem, a relevância do parâmetro t foi verificada especialmente na obtenção do mapa de cores.

Com a depuração dos dados, as taxas de acerto do classificador superaram as taxas com todo o conjunto de dados em todas as faixas de potencial eletroquímico analisadas. O destaque maior aconteceu na faixa de alto potencial, obtendo a alta taxa de 94% de acertos em conjunto com o classificador Bayes. Ressalta-se que nesta faixa de potenciais,

podem ocorrer a ruptura do filme de passivação dos aços inoxidáveis e a corrente sobe abruptamente, sendo uma forma válida de se medir a resistência à corrosão localizada das ligas, mas este aspecto não foi considerado neste trabalho, mas tão somente a estrutura dos dados originados de dois tipos de aços.

Adicionalmente, sinais provenientes de espectroscopia de impedância eletroquímica do aço inoxidável superdúplex UNS S32750 em meio corrosivo, foram também estudados. Esta impedância é uma variável complexa, e graficamente apresentada no modo de Bode (logaritmo do módulo e fase *versus* logaritmo da frequência) ou de Nyquist (parte imaginária *versus* a real). A possibilidade de trabalhar separada e conjuntamente as informações referente ao módulo e à fase de cada sinal permitiram gerar conclusões complementares que foram úteis na depuração desses sinais, introduzindo um procedimento confiável para a validação dos dados produzidos.

De modo acessório, testou-se a validade da lei de Benford no pré-processamento de dados eletroquímicos na base 10. Para as curvas de polarização foi encontrada a máxima diferença de $4,4\%$ para o aço 304 e $4,59\%$ para o aço 316 em relação ao primeiro dígito significativo. Como indicado na seção correspondente, tratam-se da diferença máxima, em percentual, entre as probabilidades dos primeiros dígitos significativos da distribuição dos dados em questão e a distribuição teórica de Benford. Para o aço 327, referente aos dados de espectroscopia de impedância eletroquímica, a máxima diferença foi de $9,65\%$ para o primeiro dígito e de $1,95\%$ para o segundo.

Da análise envolvendo o parâmetro de escala α, concluiu-se que, na presença de um conjunto de dados com possíveis ruídos, o valor para este parâmetro utilizado para definir a similaridade entre os pontos de dados é ainda bastante sensível quanto à disposição dos dados. Ao dispor de dados supostamente ruidosos (de outra forma, esparsos), uma vizinhança maior parece ser mais apropriada. É inegável a necessidade de se aprofundar no estudo do efeito do parâmetro α dos mapas de difusão.

Ao estudar a evolução dos mapas de difusão tomando o parâmetro α como se tratasse de uma variável temporal, deixando-o variar lentamente em um extenso intervalo, e considerando uma medida desses mapeamentos pela função M, foi possível caracterizar o efeito topológico de α no mapeamento. Este entendimento não havia sido relatado no trabalho de Singer et al. (2007).

Além disto, foram analisados perfis simulados segundo uma equação e com a adição de um ruido controlado. Este teste objetivou ter maior controle dos dados e permitir a análise de grupos mais ou menos intrincados. Nos perfis simulados foi possível observar que mais autovalores contribuem para a expressão do $M(\alpha)$ quando os dois *clusters* estavam mais próximos. Com grupos bem definidos, o segundo maior autovalor sobressaiu em relação aos demais e, com isso, dominou a variável $M(\alpha)$ afetando as informações do mapeamento. Por outro lado, isso pode ter acontecido devido à unicidade do parâmetro de mudança na expressão que gerou os perfis simulados.

O mapeamento obtido com os perfis simulados apresentou três faixas com características distintas. Na primeira, um aspecto desordenado provavelmente devido ao valor abaixo do limite numérico viável para a aplicação do algoritmo. A segunda, uma região de mudança de percepção da técnica principalmente relacionada à presença de *outliers* e de grupos distintos. A terceira faixa correspondeu a uma região estável e de pouca mudança na disposição dos dados mapeados. Para os dados simulados, cada faixa foi facilmente identificada no gráfico $M(\alpha)$ proposto, assim, a aplicação deste procedimento de análise para a escolha do parâmetro de escala é indicada. A sugestão deste trabalho é que a escolha do parâmetro seja feita no início da terceira faixa, para os dados aqui simulados.

Em relação à comparação da análise aqui introduzida com o que foi proposto por Singer et al. (2007) e presente em vários trabalhos na literatura, conclui-se que o procedimento apresentado de construção de $M(\alpha)$ e sua visualização dinâmica (quando possível) é interessante, pois permite a escolha de uma região segura para a escolha do parâmetro do núcleo gaussiano e fornece informações relevantes sobre a estrutura dos dados envolvidos. Em comparação com o método proposto por Shan e Daubechies (2022), no que lhe concerne, o procedimento apontado nesta tese mostrou-se coerente com os resultados encontrados aplicando o método dos autores para a escolha do parâmetro para o caso em estudo e isto indica que a ferramenta é confiável. Na ausência de conhecimento suficiente dos dados disponíveis e da natureza do processo experimental que os gerou, as rotinas aqui descritas podem ser úteis antes da aplicação da técnica de mapas de difusão.

REFERÊNCIAS

AHER, S. D.; LOBO, M. Data mining in educational system using weka. *International Journal of Computer Applications*, v. 3, p. 20–25, 2011.

ALLAH, F. A.; GROSKY, W. I.; ABOUTAJDINE, D. Document clustering based on diffusion maps and a comparison of the k-means performances in various spaces. In: *2008 Symposium on Computers and Communications*. Marrocos: IEEE, 2008.

ANGERER, P.; HAGHVERDI, L.; BüTTNER, M.; THEIS, F. J.; MARR, C.; BUETTNER, F. Destiny: diffusion maps for large-scale single-cell data in r. *Bioinformatics*, v. 32, n. 8, p. 1241–1243, 2015.

BAH, B. *Diffusion Maps: Analysis and Applications*. Dissertação (Mestrado) — University of Oxford, 2008.

BARKAN, O.; WEILL, J.; WOLF, L.; ARONOWITZ, H. Fast high dimensional vector multiplication face recognition. In: *International Conference on Computer Vision (ICCV)*. Australia: IEEE, 2013. p. 1960–1967.

BELKIN, M. *Problems of Learnings on Manifolds*. Tese (Doutorado) — The University of Chicago, Illinois, 2003.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, v. 15, n. 6, p. 1373–1396, 2003.

BENFORD, F. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, v. 78, n. 4, p. 551–572, 1938.

BENGIO, Y.; DELALLEAU, O.; ROUX, N. L.; PAIEMENT, J.-F.; VINCENT, P.; OUIMET, M. Spectral dimensionality reduction. In: *Feature Extraction*. Heidelberg: Springer Berlin Heidelberg, 2006. p. 519–550.

BERTHOLD, M. *Intelligent data analysis: an introduction*. Berlin: Springer, 2007.

BORG, P. J. F. G. I. *Modern Multidimensional Scaling*. New York: Springer-Verlag GmbH, 2007.

BRANDT, S. *Data Analysis*. Switzerland: Springer International Publishing, 2014.

CHEN, Y. F.; LIU, S.-Y.; LIU, M.; MILLER, J.; HOW, J. P. Motion planning with diffusion maps. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Daejeon: IEEE, 2016. p. 1423–1430.

COIFMAN, R. R.; LAFON, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, v. 21, n. 1, p. 5–30, 2006.

COSTA, J. I. de F. *Desenvolvimento de metodologias contabilométricas aplicadas a auditoria contábil digital: uma proposta de análise da lei de Newcomb-Benford para os Tribunais de Contas*. 447 p. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2010.

COX, M. A. A.; COX, T. F. Multidimensional scaling. In: *Handbook of Data Visualization*. Berlin, Heidelberg: Springer, 2008. p. 315–347.

DIAS, M. S. *Regressão construtiva em variedades implícitas*. Dissertação (Mestrado) — Pontíficia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2012.

DOUGLAS CARROLL, J.; ARABIE, P. Multidimensional scaling. In: BIRNBAUM, M. H. (Ed.). *Measurement, Judgment and Decision Making*. San Diego: Academic Press, 1998, (Handbook of Perception and Cognition (Second Edition)). p. 179–250.

FABBRI, R.; BASTOS, I. N.; MOURA NETO, F. D.; LOPES, F. J.; GONÇALVES, W. N.; BRUNO, O. M. Multi-q pattern classification of polarization curves. *Physica A: Statistical Mechanics and its Applications*, v. 395, p. 332–339, 2014.

GENTIL, V. *Corrosão*. Rio de Janeiro: LTC, 2011.

GHOJOGH, B.; GHODSI, A.; KARRAY, F.; CROWLEY, M. *Multidimensional Scaling, Sammon Mapping, and Isomap: Tutorial and Survey*. [S.l.]: arXiv:2009.08136, 2020.

HAGHVERDI, L.; BUETTNER, F.; THEIS, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, v. 31, n. 18, p. 2989–2998, 2015.

HILL, T. P. A statistical derivation of the significant-digit law. *Statistical Science*, v. 10, n. 4, 1995.

HOUT, M. C.; PAPESH, M. H.; GOLDINGER, S. D. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, v. 4, n. 1, p. 93–103, 2012.

IHLER, A. *Nonlinear Manifold Learning*. 2003. MIT 6.454 Summary.

JOLLIFFE, I. T. *Principal component analysis*. New York: Springer, 2002.

KELLER, Y.; COIFMAN, R. R.; LAFON, S.; ZUCKER, S. W. Audio-visual group recognition using diffusion maps. *IEEE Transactions on Signal Processing*, v. 58, n. 1, p. 403–413, 2010.

KRUSKAL, J.; WISH, M. *Multidimensional Scaling*. [S.l.]: SAGE Publications, 1978. (Quantitative Applications in the Social Sciences).

LA PORTE, J. de; HERBST, B. M.; W.HEREMAN; WALT, S. J. van der. An introduction to diffusion maps. In: *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*. Cape Town: IAPR, 2008. p. 15–25.

LAFON, S. S. *Diffusion Maps and Geometric Harmonics*. Tese (Doutorado) — Yale University, 2004.

LEE, J. A.; VERLEYSEN, M. *Nonlinear Dimensionality Reduction*. New York: Springer-Verlag GmbH, 2007.

LUO, D.; CHEN, H.; YU, H.; SUN, Y. A novel approach for classification of chinese herbal medicines using diffusion maps. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 29, n. 01, p. 1550003, 2015.

MOURA NETO, F.; SOUZA, P.; MAGALHÃES, M. de. Determining baseline profile by diffusion maps. *European Journal of Operational Research*, v. 279, n. 1, p. 107–123, 2019.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, v. 4, p. 39–40, 1881.

NIGRINI, M. J.; MITTERMAIER, L. J. The use of benford's law as an aid in analytical procedures. *Auditing: A Journal of Practice and Theory*, v. 16, n. 2, p. 52–67, 1997.

RAMPAZZO, L. *Metodologia científica: para alunos dos cursos de graduação e pós-graduação*. São Paulo: Loyola, 2013.

ROBINSON, A. *Escrita*. Porto Alegre: L&PM Pocket, 2018.

ROWEIS, S. T. Nonlinear dimensionality reduction by locally linear embedding. *Science*, v. 290, n. 5500, p. 2323–2326, 2000.

SALHOV, M.; BERMANIS, A.; WOLF, G.; AVERBUCH, A. Approximately-isometric diffusion maps. *Applied and Computational Harmonic Analysis*, v. 38, n. 3, p. 399–419, 2015.

SAUL, L. K.; ROWEIS, S. T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, v. 70, p. 119–155, 2003.

SCHöLKOPF, B.; SMOLA, A.; MüLLER, K.-R. Kernel principal component analysis. In: *Lecture Notes in Computer Science*. [S.l.]: Springer Berlin Heidelberg, 1997. p. 583–588.

SHALIZI, C. Nonlinear dimensionality reduction i: Local linear embedding. *Disponível em https://www.stat.cmu.edu/ cshalizi/350/lectures/14/lecture-14.pdf*, 2009.

SHAN, S.; DAUBECHIES, I. *Diffusion Maps : Using the Semigroup Property for Parameter Tuning*. [S.l.]: arXiv:2203.02867 [stat.ML], 2022.

SHEVCHIK, S.; ZANOLI, S.; SAEIDI, F.; MEYLAN, B.; FLüCK, G.; WASMER, K. Monitoring of friction-related failures using diffusion maps of acoustic time series. *Mechanical Systems and Signal Processing*, v. 148, p. 107–172, 2021.

SINGER, A.; ERBAN, R.; KEVREKIDIS, I. G.; COIFMAN, R. R. Detecting the slow manifold by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, p. 16090–16095, 2007.

SINGER, A.; WU, H.-T. Orientability and diffusion maps. *Applied and Computational Harmonic Analysis*, v. 31, n. 1, p. 44–58, 2011.

STERN, M.; GEARY, A. L. Electrochemical polarization. *Journal of The Electrochemical Society*, v. 104, n. 1, p. 56, 1957.

SUN, Y.-T.; TAN, X.; LEI, L.-L.; LI, J.; JIANG, Y.-M. Revisiting the effect of molybdenum on pitting resistance of stainless steels. *Tungsten*, v. 3, n. 3, p. 329–337, 2021.

TALMON, R.; COHEN, I.; GANNOT, S. Single-channel transient interference suppression with diffusion maps. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 21, n. 1, p. 132–144, 2013.

TENENBAUM, J. B. A global geometric framework for nonlinear dimensionality reduction. *Science*, v. 290, n. 5500, p. 2319–2323, 2000.

TORGERSON, W. S. Multidimensional scaling: I. theory and method. *Psychometrika 17*, 1952.

WANG, G.; LIU, J.; LI, Y.; SHANG, L. Fault detection based on diffusion maps and k nearest neighbor diffusion distance of feature space. *Journal of Chemical Engineering of Japan*, v. 48, n. 9, p. 756–765, 2015.

WANG, J. Diffusion maps. In: *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Berlin: Springer Berlin Heidelberg, 2012. p. 267–298.

WITTEN, I. H. *Data mining: practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann, 2011.

XU, R.; DAMELIN, S.; NADLER, B.; WUNSCH, D. C. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. *Artificial Intelligence in Medicine*, v. 48, n. 2-3, p. 91–98, 2010.

YANG, J.; WARD, M.; RUNDENSTEINER, E.; HUANG, S. *Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets*. Worcester, MA: The Eurographics Association, 2003.

YANG, W.; NI, R.-C.; HUA, H.-Z.; POURBAIX, A. The behavior of chromium and molybdenum in the propagation process of localized corrosion of steels. *Corrosion Science*, v. 24, n. 8, p. 691–707, 1984.

ZHANG, H.; ALBIN, S. Detecting outliers in complex profiles using a χ^2 control chart method. *IIE Transactions*, v. 41, n. 4, p. 335–345, 2009.

ANEXO A – Algoritmo PCA

Seja $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ o conjunto de dados de entrada que, por suposição, representa uma amostra de tamanho n, com n exemplares de um vetor aleatório, denotados por $\mathbf{x}_j \in \mathbb{R}$, com $j = 1, 2, \ldots, n$, e cujas d entradas são denotadas por x_{ij}, com $i = 1, 2, \ldots, d$ e $j = 1, 2, \ldots, n$. Na forma de matriz:

$$X = \begin{bmatrix} | & | & \vdots & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \vdots & \mathbf{x}_n \\ | & | & \vdots & | \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dn} \end{bmatrix} \tag{A.1}$$

Como o objetivo da técnica passa pelo cálculo do espectro dos autovalores e autovetores da matriz de covariância e esta, por sua vez, é dada em função da média amostral, é útil que a média dos dados em cada coordenada (média amostral, se visto como vetores aleatórios) seja nula. Isto é:

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j = 0. \tag{A.2}$$

De outra forma, a soma de cada linha da matriz X deve ser nula. Não há garantia, contudo, que a soma das colunas (soma das coordenadas) de cada vetor seja nula.

Para operacionalizar tal desejo, faz-se um ajuste na matriz X calculando a média de cada componente dos vetores de dados e subtraindo essa média de cada entrada da matriz referente a linha correspondente. Isto é:

$$X_{\text{ajuste}} = A = \begin{bmatrix} (x_{11} - \frac{1}{n}\sum_{k=1}^{n} x_{1k}) & (x_{12} - \frac{1}{n}\sum_{k=1}^{n} x_{1k}) & \cdots & (x_{1n} - \frac{1}{n}\sum_{k=1}^{n} x_{1k}) \\ (x_{21} - \frac{1}{n}\sum_{k=1}^{n} x_{2k}) & (x_{22} - \frac{1}{n}\sum_{k=1}^{n} x_{2k}) & \cdots & (x_{2n} - \frac{1}{n}\sum_{k=1}^{n} x_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{d1} - \frac{1}{n}\sum_{k=1}^{n} x_{dk}) & (x_{d2} - \frac{1}{n}\sum_{k=1}^{n} x_{dk}) & \cdots & (x_{dn} - \frac{1}{n}\sum_{k=1}^{n} x_{dk}) \end{bmatrix} \tag{A.3}$$

Deste modo, a matriz de covariância amostral que exibe as covariâncias entre quaisquer duas características, isto é, entre duas linhas de A será:

$$\hat{\text{Cov}}(A) = R = \frac{AA^{\intercal}}{n}. \tag{A.4}$$

Suponha agora os n dados ajustados (colunas de A — $\mathbf{a}_j$) em uma reta pela origem na direção de um vetor unitário $\mathbf{v}$. Como a projeção de cada vetor $\mathbf{a}_j$ sobre $\mathbf{v}$ tem a forma:

$$\text{proj}_{\mathbf{v}}\mathbf{a}_j = \frac{\langle \mathbf{v}, \mathbf{a}_j \rangle}{\|\mathbf{v}\|^2}\mathbf{v} \tag{A.5}$$

e $\mathbf{v}$ é unitário, logo:

$$\text{proj}_{\mathbf{v}}\mathbf{a}_j = \langle \mathbf{v}, \mathbf{a}_j \rangle \mathbf{v}, \quad j = 1, 2, \ldots, n. \tag{A.6}$$

Ao usar os vetores projeção em lugar dos vetores de dados, os erros têm tamanho:

$$\|\mathbf{a}_j - \langle \mathbf{v}, \mathbf{a}_j \rangle \mathbf{a}\|^2 \tag{A.7}$$

Fazendo $\zeta_j = \langle \mathbf{v}, \mathbf{a}_j \rangle$,

$$\begin{aligned}
\|\mathbf{a}_j - \zeta_j \mathbf{v}\|^2 &= \langle \mathbf{a}_j - \zeta_j \mathbf{v}, \mathbf{a}_j - \zeta_j \mathbf{v} \rangle \\
&= \langle \mathbf{a}_j, \mathbf{a}_j \rangle - \langle \mathbf{a}_j, \zeta_j \mathbf{v} \rangle - \langle \mathbf{a}_j, \zeta_j \mathbf{v} \rangle + \langle \zeta_j \mathbf{v}, \zeta_j \mathbf{v} \rangle \\
&= \|\mathbf{a}_j\|^2 - 2\langle \mathbf{a}_j, \zeta_j \mathbf{v} \rangle + \zeta_j^2 \|\mathbf{v}\|^2 \\
&= \|\mathbf{a}_j\|^2 - 2\zeta_j \langle \mathbf{a}_j, \mathbf{v} \rangle + \zeta_j^2 \\
&= \|\mathbf{a}_j\|^2 - 2\zeta_j \cdot \zeta_j + \zeta_j^2 \\
&= \|\mathbf{a}_j\|^2 - 2\zeta_j^2 + \zeta_j^2 \\
&= \|\mathbf{a}_j\|^2 - \zeta_j^2 \\
&= \|\mathbf{a}_j\|^2 - \langle \mathbf{v}, \mathbf{a}_j \rangle^2
\end{aligned} \tag{A.8}$$

Somando todos os n erros cometidos, tem-se a soma residual de quadrados (*residual sum of squares* — RSS):

$$\text{RSS}(\mathbf{v}) = \sum_{j=1}^{n} (\|\mathbf{a}_j\|^2 - \langle \mathbf{v}, \mathbf{a}_j \rangle^2) = \sum_{j=1}^{n} \|\mathbf{a}_j\|^2 - \sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{a}_j \rangle^2 \tag{A.9}$$

Quer-se conhecer a direção $\mathbf{v}$ tal que a soma destes resíduos seja a menor possível. Na verdade, em geral, é desejado não só projetar em apenas uma direção, mas em múltiplas componentes principais. O mesmo resultado pode, no entanto, ser estendido para m direções com $\mathbf{v}$ como uma matriz $V_{m \times d}$, ao invés de $1 \times d$. Para isso, temos que maximizar $\sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{a}_j \rangle^2$ ou, de forma conveniente, $\frac{1}{n} \sum_{j=1}^{n} \langle \mathbf{v}, \mathbf{a}_j \rangle^2$ uma vez que $\mathbf{v}$ não depende de n.

Um resultado útil é que, para uma variável aleatória, a média(esperança) de seu quadrado é o quadrado da média mais a variância.

$$E(X^2) = E^2(X) + \hat{\text{Var}}(X) \tag{A.10}$$

Sabendo que um resultado análogo também é válido para os estimadores, pode-se escrever:

$$\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle^2 = \left(\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle\right)^2 + \hat{\mathrm{Var}}(\langle \mathbf{v}, \mathbf{a}_j\rangle). \tag{A.11}$$

Como a soma dos vetores representados pelas colunas de A também é nula, a média das projeções também é nula e, assim, $\left(\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle\right)^2 = 0$. De fato, tem-se:

$$\begin{aligned}\left(\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle\right)^2 &= \frac{1}{n^2}\left(\langle \mathbf{v}, \mathbf{a}_1\rangle + \langle \mathbf{v}, \mathbf{a}_2\rangle + \ldots + \langle \mathbf{v}, \mathbf{a}_n\rangle\right)^2 \\ &= \frac{1}{n^2}\left(\langle \mathbf{v}, \mathbf{a}_1 + \mathbf{a}_2 + \ldots + \mathbf{a}_n\rangle\right)^2 \\ &= \frac{1}{n^2}\left(\langle \mathbf{v}, \mathbf{0}\rangle\right)^2 \\ &= \frac{1}{n^2}(0)^2 \\ &= 0\end{aligned} \tag{A.12}$$

Logo,

$$\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle^2 = \mathrm{Var}(\langle \mathbf{v}, \mathbf{a}_j\rangle). \tag{A.13}$$

Ou seja, o objetivo é determinar a direção de projeção $\mathbf{v}$ na qual os dados projetados $\langle \mathbf{v}, \mathbf{a}_j\rangle$ apresentam maior variância amostral. Mas,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\langle \mathbf{v}, \mathbf{a}_j\rangle^2 &= \frac{1}{n}\|\mathbf{v}^\intercal A\|^2 \\ &= \frac{1}{n}\langle \mathbf{v}^\intercal A, \mathbf{v}^\intercal A\rangle \\ &= \frac{1}{n}(\mathbf{v}^\intercal A)(\mathbf{v}^\intercal A)^\intercal \\ &= \frac{1}{n}\mathbf{v}^\intercal A A^\intercal \mathbf{v} \\ &= \mathbf{v}^\intercal \frac{AA^\intercal}{n}\mathbf{v} \\ &= \mathbf{v}^\intercal R \mathbf{v}\end{aligned} \tag{A.14}$$

Logo, quer-se encontrar $\mathbf{v}$ que maximize $\mathbf{v}^\intercal R\mathbf{v}$ sujeito à condição $\mathbf{v}^\intercal\mathbf{v} = 1$. Ou seja,

$$\arg\max_{\mathbf{v}} \mathbf{v}^\intercal R\mathbf{v} \tag{A.15}$$

com $\mathbf{v}^\intercal\mathbf{v} = 1$.

Seja a função $f(\mathbf{v}) = \mathbf{v}^\intercal R\mathbf{v}$ ao qual quer-se maximizar. Também tem-se que

$g(\mathbf{v}) = \mathbf{v}^\intercal\mathbf{v} = 1$ que, de forma conveniente, pode ser escrita como $g(\mathbf{v}) - 1 = 0$. Adiciona-se o multiplicador de Lagrange ς e considera-se a função $F(\mathbf{v}, \varsigma) = f(\mathbf{v}) - \varsigma(g(\mathbf{v}) - 1)$. Diferenciando em relação a ambos os argumentos e igualando a zero:

$$\begin{aligned} &\frac{\partial F}{\partial \mathbf{v}} = 2R\mathbf{v} - 2\varsigma\mathbf{v} = 0 \\ &R\mathbf{v} = \varsigma\mathbf{v} \end{aligned} \tag{A.16}$$

e

$$\begin{aligned} &\frac{\partial F}{\partial \varsigma} = -(g(\mathbf{v}) - 1) = 0 \\ &g(\mathbf{v}) = 1 \end{aligned} \tag{A.17}$$

Assim, o vetor $\mathbf{v}$ desejado é um autovetor da matriz de covariância R, e a maximização deste está associado aos maiores autovalores ς. Como R é uma matriz positiva, todos seus autovalores são reais e não negativos com seus d diferentes autovetores ortogonais a qualquer outro. Ainda, o maior valor da função f ocorre para $\mathbf{v} = \mathbf{1}$, associado ao maior autovalor, ς_1. De fato, $f(\mathbf{1}) = \mathbf{1}^\intercal R\mathbf{1} = \mathbf{1}^\intercal\varsigma_1\mathbf{1} = \varsigma_1\mathbf{1}^\intercal\mathbf{1} = \varsigma_1$.

Em resumo:

Algoritmo PCA

ENTRADA: Conjunto de dados de alta dimensão $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ e a dimensão m pretendida para a redução.

1. Calcule a média de cada componente dos vetores de dados $\mathbf{x}_i$, com $i = 1, 2, \ldots, n$. Isto é equivalente a calcular a média de cada linha da matriz X onde os vetores de dados são suas colunas.

2. Calcule a matriz A subtraindo de cada entrada x_{ij} da matriz X a média da linha i (Equação A.3). Isto vai simplificar o cálculo da matriz de covariância, que passa a ser dada por $\hat{\text{Cov}}(A) = R = \dfrac{AA^\intercal}{n}$.

3. Calcule os m autovetores da matriz de covariância R associados aos m maiores autovalores e disponha-os em uma matriz B com cada um ordenado em uma coluna na ordem crescente. A matriz B será do tipo $d \times m$ onde a primeira coluna conterá o autovetor que corresponde ao autovalor dominante, a segunda coluna o autovetor que corresponde ao segundo autovalor dominante, etc.

4. Mapeie os dados para o novo espaço m-dimensional, fazendo $Z = B^\intercal A$. Com isso, Z é a matriz $m \times n$ que contém em suas colunas as novas coordenadas dos dados no novo espaço. De igual forma:

$$Z = \begin{bmatrix} | & | & \vdots & | \\ \mathbf{z}_1 & \mathbf{z}_2 & \vdots & \mathbf{z}_n \\ | & | & \vdots & | \end{bmatrix}, \quad \text{com} \quad \mathbf{z}_i = \begin{bmatrix} \sum_{l=1}^{d} \delta_1(l) a_{li} \\ \sum_{l=1}^{d} \delta_2(l) a_{li} \\ \vdots \\ \sum_{l=1}^{d} \delta_m(l) a_{li} \end{bmatrix}$$

onde $\delta_1(l)$ indica o l-ésimo elemento do primeiro autovetor de R associado ao maior autovalor, $\delta_2(l)$ o l-ésimo elemento do segundo autovetor de R associado ao segundo maior autovalor, e assim por diante.

SAÍDA: Conjunto de dados mergulhados em um espaço de menor dimensão $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^m$.

ANEXO B – Algoritmo LLE

A imersão localmente linear é uma técnica para encontrar coordenadas globais de baixa dimensão quando os dados se encontram inseridos, ou muito próximos, a uma variedade embutido em um espaço de alta dimensão. Com esse fim, utilizam-se dos q vizinhos próximos a dados específicos para aproximação da variedade por um espaço euclidiano local, onde as propriedades das operações vetoriais são válidas.

O procedimento LLE tem três etapas: constrói-se uma vizinhança para cada ponto de dados, encontram-se os pesos para aproximar linearmente os dados naquela vizinhança e, finalmente, obtém-se as coordenadas de baixa dimensão melhor reconstruídas por esses pesos (SHALIZI, 2009). Em seguida, estas coordenadas de baixa dimensão são então retornadas. Adiante, é descrito em detalhes o procedimento teórico e prático de cada etapa.

Etapa 1: Encontrando os vizinhos

Na etapa 1, deve-se estabelecer uma vizinhança para cada ponto de dados e fazê-lo de uma forma que se conforme ou se adapte aos dados. Segundo Shalizi (2009), não é estritamente necessário escolher usar q vizinhos mais próximos. Resultados semelhantes são obtidos para diferentes valores de q, com $q \geq m + 1$, com m o número de dimensões desejado para o espaço de imersão.

Para encontrar os q vizinhos mais próximos de cada ponto, calcula-se as distâncias entre todos os pares de pontos. As vizinhanças dependem apenas dessas distâncias e não dos pontos propriamente ditos. Na prática, basta encontrar as q menores entradas em cada linha da matriz de distâncias.

Com o *Wolfram Mathematica 12*, pode-se usar a função *DistanceMatrix*[]. Ela retornará uma matriz $n \times n$ onde cada entrada ij contém a distância entre o vetor $\mathbf{x}_i$ e o vetor $\mathbf{x}_j$. Busca-se nesta matriz o índice em cada linha referente a menor das distâncias. Com a função *Ordering*[*list*] consegue-se encontrar os índices referentes as $q + 1$ menores entradas em cada linha. Lembrando que deve-se excluir a entrada zero em cada linha correspondente à distância de cada vetor de dados a ele mesmo.

Etapa 2: Encontrando pesos

A etapa seguinte leva em conta que toda variedade é localmente linear. Assim sendo, é suposto que esta variedade seja exatamente linear em torno de um vetor de dados $\mathbf{x}_i$, ou seja, que ele e seus vizinhos pertençam a um subespaço linear m-dimensional. Como $m + 1$ vetores podem gerar um subespaço m-dimensional, $\mathbf{x}_i$ pode ser escrito como

combinação linear dos vizinhos, ou seja,

$$\mathbf{x}_i = \sum_j e_{ij}\mathbf{x}_j \tag{B.1}$$

onde e_{ij} são pesos apropriados que, idealmente, devem reconstruir $\mathbf{x}_i$ tanto no espaço original de alta dimensão quanto no subespaço de imersão de baixa dimensão. De acordo com Shalizi (2009), são os pesos em torno de um determinado ponto que caracterizam a aparência da variedade, desde que a vizinhança seja pequena o suficiente em comparação com a curvatura. Descobrir os pesos nos dá a mesma informação que encontrar o espaço tangente.

Uma restrição imposta aos pesos é que $\sum_{j=1}^{n} e_{ij} = 1$. Isto, geometricamente, garante a invariância sob translação. Se um vetor $\mathbf{c}$ qualquer é adicionado a $\mathbf{x}_i$ e a todos os seus vizinhos, nada acontece com a função que deseja-se minimizar,

$$\begin{aligned}\mathbf{x}_i + \mathbf{c} - \sum_j e_{ij}(\mathbf{x}_j + \mathbf{c}) &= \mathbf{x}_i + \mathbf{c} - \left(\sum_j e_{ij}\mathbf{x}_j\right) - \mathbf{c} \\ &= \mathbf{x}_i - \sum_j e_{ij}\mathbf{x}_j\end{aligned} \tag{B.2}$$

Por outro lado, se a soma dos pesos é um, $(E)_{ij} = e_{ij}$ é uma matriz de transição estocástica ou matriz de *Markov*.

Em resumo, para cada $\mathbf{x}_i$, deseja-se encontrar os pesos e_{ij} que minimizam

$$\mathrm{RSS}_i(\mathbf{e}) = \|\mathbf{x}_i - \sum_j e_{ij}\mathbf{x}_j\|^2 \tag{B.3}$$

onde $e_{ij} = 0$, ao menos que $\mathbf{x}_j$ seja um dos q-vizinhos mais próximos de $\mathbf{x}_i$ e, para cada i, $\sum_{j=1}^{n} e_{ij} = 1$. Como RSS_i é invariante a adição de um vetor c arbitrário, pode-se definir $\mathbf{c} = -\mathbf{x}_i$, centralizando os vetores no ponto focal $\mathbf{x}_i$:

$$\begin{aligned}\mathrm{RSS}_i(\mathbf{e}) &= \|\mathbf{x}_i + \mathbf{c} - \sum_j e_{ij}(\mathbf{x}_j + \mathbf{c})\|^2 \\ &= \|\sum_j e_{ij}(\mathbf{x}_j - \mathbf{x}_i)\|^2 \\ &= \|\sum_j e_{ij}\mathbf{h}_j\|^2\end{aligned} \tag{B.4}$$

com $\mathbf{h}_j = \mathbf{x}_j - \mathbf{x}_i$. Se $H_{q\times d}$ é a matriz com as entradas $\mathbf{h}_j$ correspondentes e $\mathbf{e}_i$ um vetor $q \times 1$ para cada linha de E, a soma na equação B.4 na forma matricial é $\mathbf{e}_i^\intercal H$. Como $\mathbf{e}_i^\intercal H$ é um vetor linha,

$$\mathrm{RSS}_i(\mathbf{e}) = \mathbf{e}_i^\intercal H H^\intercal \mathbf{e}_i \tag{B.5}$$

com $HH^\intercal$ uma matriz $q \times q$ com todos os produtos internos dos vizinhos. Fazendo $H^* = HH^\intercal$ e lembrando que H^* depende da escolha do ponto focal $\mathbf{x}_i$,

$$\mathrm{RSS}_i(\mathbf{e}) = \mathbf{e}_i^\intercal H_i^* \mathbf{e}_i \tag{B.6}$$

Ao minimizar $\mathrm{RSS}_i(\mathbf{e})$ com a restrição $\sum_{j=1}^n e_{ij} = 1$, introduz-se o multiplicador de Lagrange ϱ. Assim, a restrição tem a forma $\mathbf{1}^\intercal \mathbf{e}_i = 1$, ou $\mathbf{1}^\intercal \mathbf{e}_i - 1 = 0$, e escrevendo na forma lagrangiana,

$$\mathcal{L}(\mathbf{e}_i, \varrho) = \mathbf{e}_i^\intercal H_i^* \mathbf{e}_i - \varrho(\mathbf{1}^\intercal \mathbf{e}_i - 1) \tag{B.7}$$

Tomando derivadas parciais e lembrando que H_i^* é simétrica,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{e}_i} = 2H_i^* \mathbf{e}_i - \varrho \mathbf{1} = 0 \tag{B.8}$$

ou

$$H_i^* \mathbf{e}_i = \frac{\varrho}{2}\mathbf{1} \tag{B.9}$$

Se a matriz H_i^* é invertível,

$$\mathbf{e}_i = \frac{\varrho}{2}(H_i^*)^{-1}\mathbf{1} \tag{B.10}$$

onde ϱ deve ser ajustado para garantir que a soma das entradas de $\mathbf{e}_i$ seja 1.

Shalizi (2009) chama a atenção para o caso que $q > d$. Se q, o número de vizinhos, for maior que d, número de características, então, em geral, o espaço gerado por q vetores distintos é o espaço inteiro. Assim, $\mathbf{x}_i$ pode ser escrito como uma combinação linear de seus q vizinhos mais próximos, contudo, esta solução não é única. De fato, se $q > d$, então não há somente uma solução para $\mathbf{x}_i = \sum_{j=1}^n e_{ij}\mathbf{x}_j$, mas, geralmente, infinitas, pois há mais incógnitas (q) do que equações (d). Neste caso, o autor apresenta uma regularização para o problema mal posto conhecida como regularização de *Tikhonov* e procede como a seguir. Ao invés de minimizar

$$\|\mathbf{x}_i - \sum_j e_{ij}\mathbf{x}_i\|^2, \tag{B.11}$$

toma-se $\varpi > 0$ e minimiza

$$\|\mathbf{x}_i - \sum_j e_{ij}\mathbf{x}_i\|^2 + \varpi \sum_j e_{ij}^2 \tag{B.12}$$

com a justificativa de que o ajuste melhora a busca pelos pesos. Com isso, a função

objetivo é modificada e passa a ter a forma

$$\mathbf{e}_i^\intercal H_i^* \mathbf{e}_i + \varpi \mathbf{e}_i^\intercal \mathbf{e}_i \tag{B.13}$$

onde ϖ determina o grau de regularização. Assim, a lagrangiana se torna

$$\mathcal{L}(\mathbf{e}_i, \varpi, \varrho) = \mathbf{e}_i^\intercal H_i^* \mathbf{e}_i + \varpi \mathbf{e}_i^\intercal \mathbf{e}_i - \varrho(\mathbf{1}^\intercal \mathbf{e}_i - 1) \tag{B.14}$$

onde, novamente, ϱ é o multiplicador de Lagrange. Derivando com respeito a $\mathbf{e}_i$ e igualando a zero,

$$\begin{aligned} 2H_i^* \mathbf{e}_i + 2\varpi \mathbf{e}_i - \varrho \mathbf{1} &= 0 \\ 2(H_i^* \mathbf{e}_i + \varpi \mathbf{e}_i) &= \varrho \mathbf{1} \\ (H_i^* + \varpi I)\mathbf{e}_i &= \frac{\varrho}{2}\mathbf{1} \\ \mathbf{e}_i &= \frac{\varrho}{2}(H_i^* + \varpi I)^{-1}\mathbf{1} \end{aligned} \tag{B.15}$$

onde, outra vez, ϱ é escolhido de modo a deixar $\mathbf{e}_i$ normalizado.

Etapa 3: Encontrando as coordenadas

Com os pesos já calculados, na terceira e última etapa do LLE, os pesos na incorporação são considerados aproximadamente iguais aos do espaço de recursos e são utilizados para comporem as coordenadas de cada dado para o mapeamento. Assim sendo, toma-se a matriz de peso E fixa e deseja-se encontrar Z que minimize:

$$\mathbf{\Phi}(Z) = \sum_i \left\| \mathbf{z}_i - \sum_{j \neq i} e_{ij} \mathbf{z}_j \right\|^2 \tag{B.16}$$

sujeito às restrições,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i = 0. \tag{B.17}$$

e

$$\frac{1}{n} Z^\intercal Z = I. \tag{B.18}$$

Enquanto que a primeira garante uma conveniência de cálculo, a segunda faz com que a matriz de covariância de Z seja a matriz identidade m-dimensional.

Como na PCA, assume-se primeiramente $m = 1$ (mapeamento unidimensional) e o caso geral é feito de modo similar. Desta forma, $\mathbf{z}_i$ é um número real, z_i, e Z se reduz a

um vetor coluna $n \times 1$. A função objetivo torna-se:

$$\begin{aligned}
\mathbf{\Phi}(Z) &= \sum_{i=1}^{n} \left(z_i - \sum_j e_{ij} z_j \right)^2 \\
&= \sum_{i=1}^{n} z_i^2 - z_i \left(\sum_j e_{ij} z_j \right) - \left(\sum_j e_{ij} z_j \right) z_i + \left(\sum_j e_{ij} z_j \right)^2 \\
&= Z^\intercal Z - Z^\intercal (EZ) - (EZ)^\intercal Z + (EZ)^\intercal (EZ) \\
&= Z^\intercal (Z - EZ) - (EZ)^\intercal (Z - EZ) \\
&= (Z^\intercal - (EZ)^\intercal)(Z - EZ) \\
&= (Z^\intercal - Z^\intercal E^\intercal)(I - E)Z \\
&= Z^\intercal (I - E)^\intercal (I - E) Z
\end{aligned} \tag{B.19}$$

Definindo a matriz $\Xi_{n \times n}$ como $\Xi = (I - E)^\intercal (I - E)$,

$$\mathbf{\Phi}(Z) = Z^\intercal \Xi Z \tag{B.20}$$

Usando o multiplicador de Lagrange μ para inserir a restrição $\frac{1}{n} Z^\intercal Z = I$,

$$\mathcal{L}(Z, \mu) = Z^\intercal \Xi Z - \mu \left(\frac{1}{n} Z^\intercal Z - 1 \right) \tag{B.21}$$

Diferenciando,

$$\frac{\partial \mathcal{L}}{\partial Z} = 2\Xi Z - \frac{2\mu}{n} Z = 0 \tag{B.22}$$

ou

$$\Xi Z = \frac{\mu}{n} Z \tag{B.23}$$

Logo, Z deve ser um autovetor de Ξ. Como o objetivo é a minimização, quer-se que as autofunções fiquem com os menores autovalores, ou seja, as autofunções inferiores. Os autovalores são reais e não negativos com o menor deles iguais a zero com autofunção 1. Como essa autofunção é constante, ela deve ser descartada. Assim, a fim de obter o mapeamento 1D, basta tomar as duas autofunções inferiores e descartar a constante. Com $m > 1$, tomam-se $m + 1$ autofunções inferiores, descarta-se a autofunção constante com autovalor nulo e utilizam-se as demais como coordenadas para o mapeamento. Como as autofunções são ortogonais, a restrição de não covariância é automaticamente satisfeita.

Algoritmo LLE

ENTRADA: Conjunto de dados de alta dimensão $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, inteiro q de vizinhos locais, parâmetro ϖ para a regularização de *Thikonov* e a dimensão m pretendida para a redução.

1. Encontre, para cada dado $\mathbf{x}_i$, q vizinhos próximos. Armazene os índices destes vizinhos em uma matriz $n \times q$. Cada linha desta matriz deve conter q números que são os índices dos vetores mais próximos de $\mathbf{x}_i$.

2. Encontre a matriz peso E que minimiza a soma residual dos quadrados (RSS - *residual sum of squares*) para reconstruir cada $\mathbf{x}_i$ a partir de seus vizinhos,

   $$\mathrm{RSS}(W) = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j \neq i} e_{ij}\mathbf{x}_j\|^2$$

   onde $e_{ij} = 0$, ao menos que $\mathbf{x}_j$ seja um dos q-vizinhos mais próximos de $\mathbf{x}_i$ e, para cada i, $\sum_{j=1}^{n} e_{ij} = 1$.

   Para tal:

   2.1. Encontre os vetores $\mathbf{h}_j = \mathbf{x}_j - \mathbf{x}_i$ fazendo j igual a todos os índices dos vetores que estão na vizinhança de um dado $\mathbf{x}_i$. Armazene os $\mathbf{h}_j$ em uma matriz H_i onde cada linha contém um vetor $\mathbf{h}_j$;

   2.2. Calcule, para cada i, $H_i^* = H_i H_i^\intercal$;

   2.3. Calcule, para cada i, $\mathbf{e}_i = \frac{\varrho}{2}(H_i^* + \varpi I)^{-1}\mathbf{1}$ de modo que $\sum_{j=1}^{n} e_{ij} = 1$;

   2.4. Escreva a matriz E onde $e_{ij} = 0$, se $\mathbf{x}_j$ não é um dos q-vizinhos mais próximos de $\mathbf{x}_i$ e, e_{ij} é igual as respectivas entradas de $\mathbf{e}_i$ para os índices correspondentes, caso contrário.

3. Encontre as coordenadas Z que minimizam o erro de reconstrução usando os pesos,

   $$\mathbf{\Phi}(Z) = \sum_{i=1}^{n} \|\mathbf{z}_i - \sum_{j \neq i} e_{ij}\mathbf{z}_j\|^2$$

   sujeito às restrições $\sum_{i=1}^{n} \mathbf{z}_{ij} = 0$, para cada j, e também $Z^\intercal Z = I$.

   Para tal:

   3.1. Defina a matriz $\Xi = (I - E)^\intercal(I - E)$.

3.2. Calcule $m+1$ autovetores de Ξ correspondentes aos $m+1$ menores autovalores.

3.3. Preencha as colunas de Z, a partir da primeira coluna, com os autovetores de Ξ correspondentes aos $m+1$ menores autovalores excluindo o primeiro autovetor constante. O autovetor correspondente ao segundo menor autovalor de Ξ entra na primeira coluna de Z, o autovetor correspondente ao terceiro menor autovalor de Ξ entra na segunda coluna de Z, e assim por diante.

SAÍDA: Conjunto de dados mergulhados em um espaço de menor dimensão $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^m$.

ANEXO C – Função custo do LE

A função custo do LE como definida na equação 2, pode ser escrita na formal matricial $f(Z) = tr(ZLZ^\intercal)$. Para tal, basta observar que para um mapeamento m-dimensional:

$$\begin{aligned}
f(Z) &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\mathbf{z}_i - \mathbf{z}_j)^\intercal(\mathbf{z}_i - \mathbf{z}_j) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[w_{ij}\mathbf{z}_i^\intercal\mathbf{z}_i - w_{ij}\mathbf{z}_i^\intercal\mathbf{z}_j - w_{ij}\mathbf{z}_j^\intercal\mathbf{z}_i + w_{ij}\mathbf{z}_j^\intercal\mathbf{z}_j\right] \\
&= \frac{1}{2}\left[\sum_{i=1}^{n} g_{ii}\mathbf{z}_i^\intercal\mathbf{z}_i - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbf{z}_i^\intercal\mathbf{z}_j + \sum_{j=1}^{n} g_{jj}\mathbf{z}_j^\intercal\mathbf{z}_j\right] \\
&= \frac{1}{2}\left[2\sum_{i=1}^{n} g_{ii}\mathbf{z}_i^\intercal\mathbf{z}_i - 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbf{z}_i^\intercal\mathbf{z}_j\right] \\
&= \sum_{i=1}^{n} g_{ii}\mathbf{z}_i^\intercal\mathbf{z}_i - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbf{z}_i^\intercal\mathbf{z}_j \\
&= tr(GZ^\intercal Z) - tr(WZ^\intercal Z)
\end{aligned}$$

Como o traço de uma matriz é um operador invariante sobre permutações cíclicas, tem-se:

$$\begin{aligned}
f(Z) &= tr(GZ^\intercal Z) - tr(WZ^\intercal Z) \\
&= tr(ZGZ^\intercal) - tr(ZWZ^\intercal) \\
&= tr(Z(G-W)Z^\intercal) \\
&= tr(ZLZ^\intercal)
\end{aligned} \tag{C.1}$$

ANEXO D – Distâncias de difusão e o espectro da matriz de difusão

O objetivo dessa seção é mostrar que se as coordenadas de difusão forem escolhidas de modo que sejam iguais às autofunções determinadas pelos autovalores da matriz de difusão P, a distância de difusão entre dois pontos no espaço de recurso será igual à distância euclidiana entre esses dois pontos mapeados no espaço de difusão. Para tal, essa demonstração se utiliza de três lemas e pode também ser encontrada em (LA PORTE et al., 2008).

Seja K uma matriz de similaridade obtida através de um núcleo de similaridade k, $(K)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, Y uma matriz diagonal que normaliza as linhas de K, $(Y)_{ii} = \sum_{j=1}^{n} k_{ij}$, e P a matriz de difusão dada por:

$$P = Y^{-1}K \tag{D.1}$$

Define-se a matriz K normalizada por:

$$K^* = Y^{-\frac{1}{2}}KY^{-\frac{1}{2}} \tag{D.2}$$

Com estas definições, podem-se enunciar os seguintes lemas:

Lema 1: A matriz K^*:

1. é simétrica;
2. tem os mesmos autovalores de P;
3. seus autovetores são conjugados por $Y^{-\frac{1}{2}}$ e $Y^{\frac{1}{2}}$ para obter os autovetores esquerdos e direitos de P, respectivamente.

Demonstração:

A simetria de K^* acontece, pois K é simétrica.
Multiplicando-se Y à esquerda de (D.1) e substituindo-a em (D.2), temos:

$$K^* = Y^{-\frac{1}{2}}YPY^{-\frac{1}{2}}$$
$$K^* = Y^{\frac{1}{2}}PY^{-\frac{1}{2}} \tag{D.3}$$

Isolando P em (D.3),

$$P = Y^{-\frac{1}{2}}K^*Y^{\frac{1}{2}} \tag{D.4}$$

Como K^* é simétrica, então existe um conjunto ortonormal de autovetores de K^* tal que:

$$K^* = S\Lambda S^T, \tag{D.5}$$

onde Λ é uma matriz diagonal contendo seus autovalores reais e S é uma matriz com os seus autovetores ortonormais nas colunas. Substituindo (D.5) em (D.4), tem-se:

$$P = Y^{-\frac{1}{2}} S\Lambda S^T Y^{\frac{1}{2}} \tag{D.6}$$

Como S é uma matriz ortogonal,

$$\begin{aligned} P &= Y^{-\frac{1}{2}} S\Lambda S^{-1} Y^{\frac{1}{2}} \\ &= (Y^{-\frac{1}{2}} S)\Lambda (Y^{-\frac{1}{2}} S)^{-1} \\ &= Q\Lambda Q^{-1} \end{aligned} \tag{D.7}$$

Com isso, os autovalores de K^* e P são iguais. Além disso, os autovetores direitos de P são as colunas de,

$$Q = Y^{-\frac{1}{2}} S \tag{D.8}$$

e os autovetores esquerdos são as linhas de

$$Q^{-1} = S^T Y^{\frac{1}{2}} \tag{D.9}$$

Obtemos uma equação para os autovetores de P em termos dos autovetores ϕ_l de K^*. Os autovetores direitos de P são dados por

$$\psi_l = Y^{-\frac{1}{2}} \phi_l \tag{D.10}$$

e os esquerdos por

$$\omega_l = Y^{\frac{1}{2}} \phi_l \tag{D.11}$$

Lema 2:

1. A matriz de difusão P tem a decomposição espectral

$$P = \sum_{l=1}^{n} \lambda_l \psi_l \omega_l^{\intercal} \tag{D.12}$$

com

$\psi_l = Y^{-\frac{1}{2}}\phi_l$
$\omega_l = Y^{\frac{1}{2}}\phi_l$

e ϕ_l o conjunto ortonormal de autovetores da matriz simétrica K^*.

2. Os autovetores esquerdos de P formam uma base ortonormal do sistema de coordenadas de $\mathbb{R}^n$, dado pela métrica Y^{-1}.

Demonstração:

De D.7, obtém-se a seguinte decomposição espectral:

$$P = \sum_l \lambda_l \psi_l \omega_l^\intercal \tag{D.13}$$

que expressa cada linha da matriz de difusão em termos da nova base ω_l. Neste novo sistema de coordenadas em $\mathbb{R}^n$, uma linha i de P é representada pelo ponto:

$$\mathbf{z}_i = \begin{bmatrix} \lambda_1\psi_1(i) \\ \lambda_2\psi_2(i) \\ \vdots \\ \lambda_n\psi_n(i) \end{bmatrix} \tag{D.14}$$

onde $\psi_n(i)$ é o i-ésimo componente do n-ésimo autovetor direito. Contudo, P não é simétrica, e então, o sistema de coordenadadas não será ortonormal, *i.e*,

$$\omega_l^\intercal \omega_l \neq 1 \tag{D.15}$$

ou, de forma equivalente,

$$\omega_l^\intercal I \omega_l \neq 1 \tag{D.16}$$

e

$$\omega_k^\intercal I \omega_l \neq 0 \qquad \text{para} \qquad k \neq l \tag{D.17}$$

No entanto, usando a matriz $Q = Y^{-1}$ simétrica positiva definida e a equação D.9,

$$\begin{aligned}\omega_l^\intercal Q\omega_l &= \omega_l^\intercal (Y^{-\frac{1}{2}})^\intercal (Y^{-\frac{1}{2}})\omega_l \\ &= \phi_l^\intercal \phi_l \\ &= 1 \end{aligned} \tag{D.18}$$

e, de igual forma,

$$\omega_l^\intercal Q\omega_k = 0 \qquad \text{para} \qquad l \neq k. \tag{D.19}$$

Assim sendo, os autovetores esquerdos da matriz de difusão formam um sistema de coordenadas ortonormais de $\mathbb{R}^n$ com métrica Y^{-1}, denotado por $L_2(\mathbb{R}^n, Y^{-1})$. Em resumo, para dois vetores $\mathbf{v}$ e $\mathbf{v}$' quaisquer nesse espaço, a distância entre eles é dada por:

$$\|\mathbf{v} - \mathbf{v}'\|^2 = (\mathbf{v} - \mathbf{v}')^\intercal Y^{-1}(\mathbf{v} - \mathbf{v}') \tag{D.20}$$

Lema 3: Sejam as coordenadas de difusão como definido em (17). As distâncias de difusão entre dois pontos quaisquer no espaço de recurso é igual à distância euclidiana no espaço de difusão.

Demonstração:

Em termos matemáticos, queremos mostrar que

$$\begin{aligned} D_t^2(\mathbf{x}_i, \mathbf{x}_j) &= \|p_t(\mathbf{x}_i, .) - p_t(\mathbf{x}_j, .)\|^2_{L_2(\mathbb{R}^n, Y^{-1})} \\ &= \|\mathbf{z}_i - \mathbf{z}_j\|^2_{L_2(\mathbb{R}^n, I)} \\ &= \sum_{l=1}^{n} \lambda_l^{2t}(\psi_l(i) - \psi_l(j))^2 \end{aligned} \tag{D.21}$$

Por simplicidade, assume-se que $t = 1$ com X o conjunto de dados. Então:

$$\begin{aligned} D^2(\mathbf{x}_i, \mathbf{x}_j) &= \|p_1(\mathbf{x}_i, .) - p_1(\mathbf{x}_j, .)\|^2_{L_2(\mathbb{R}^n, Y^{-1})} \\ &= \left\| \sum_l \lambda_l \psi_l(i)\omega_l^\intercal - \sum_l \lambda_l \psi_l(j)\omega_l^\intercal \right\|^2_{L_2(\mathbb{R}^n, Y^{-1})} \\ &= \left\| \sum_l \lambda_l \omega_l^\intercal (\psi_l(i) - \psi_l(j)) \right\|^2_{L_2(\mathbb{R}^n, Y^{-1})} \end{aligned}$$

$$
\begin{aligned}
&= \left\| \sum_l \lambda_l \phi_l^\intercal (\psi_l(i) - \psi_l(j)) Y^{\frac{1}{2}} \right\|^2_{L_2(\mathbb{R}^n, Y^{-1})} \\
&= \left(\sum_l \lambda_l \phi_l^\intercal (\psi_l(i) - \psi_l(j)) Y^{\frac{1}{2}} \right) Y^{-1} \left(\sum_o \lambda_o \phi_o^\intercal (\psi_o(i) - \psi_o(j)) Y^{\frac{1}{2}} \right)^\intercal \\
&= \left(\sum_l \lambda_l \phi_l^\intercal (\psi_l(i) - \psi_l(j)) Y^{\frac{1}{2}} \right) Y^{-1} \left(Y^{\frac{1}{2}} \sum_o \lambda_o \phi_o (\psi_o(i) - \psi_o(j)) \right) \\
&= \sum_l \lambda_l \phi_l^\intercal (\psi_l(i) - \psi_l(j)) \sum_o \lambda_o \phi_o (\psi_o(i) - \psi_o(j))
\end{aligned} \tag{D.22}
$$

Uma vez que $\{\phi_l\}$ é um conjunto ortonormal,

$\phi_l^\intercal \phi_o = 0 \qquad \text{para} \qquad l \neq o.$

e, finalmente,

$$
\begin{aligned}
D^2(\mathbf{x}_i, \mathbf{x}_j) &= \sum_l \lambda_l^2 (\psi_l(i) - \psi_l(j))^2 \\
&= \|\mathbf{z}_i - \mathbf{z}_j\|^2
\end{aligned} \tag{D.23}
$$

ANEXO E – Trabalhos e artigos publicados

Durante a realização do curso e da escrita desta tese, foram publicados 1(um) resumo expandido, 3(três) trabalhos completos e 1(um) artigo que abordam parte da pesquisa original deste trabalho final. Estes são aqui listados.

[1] OLIVEIRA, C. A. L. S; MOURA NETO, F. D.; BASTOS, I. N. Análise do tamanho da vizinhança no mapeamento de curvas de polarização por mapas de difusão. In: Anais da VIII Exposição de Trabalhos Acadêmicos da Região Serrana (ETARSERRA) 2020 . Disponível em:<https://www.even3.com.br/anais/adedtadrs22021/498902-analise-do-tamanho-da-vizinhanca-no-mapeamento-de-curvas-de-polarizacao-por-mapas-de-difusao/> . Acesso em: 01/08/2022.

[2] OLIVEIRA, C. A. L. S; MOURA NETO, F. D.; BASTOS, I. N. Organização eficiente de imagens digitais por mapas de difusão. In: Anais do XXIII Encontro Nacional de Modelagem Computacional e o XI Encontro de Ciência e Tecnologia de Materiais [recurso eletrônico], 28 a 30 de outubro de 2020, Palmas, TO, Brasil. ISSN: 2527-2357.

[3] OLIVEIRA, C. A. L. S; MOURA NETO, F. D.; BASTOS, I. N. Mapas de difusão na Busca de Outliers de Sinais Eletroquímicos. In: Anais do XXIII Encontro Nacional de Modelagem Computacional e o XI Encontro de Ciência e Tecnologia de Materiais [recurso eletrônico], 28 a 30 de outubro de 2020, Palmas, TO, Brasil. ISSN: 2527-2357.

[4] OLIVEIRA, C. A. L. S; MOURA NETO, F. D.; BASTOS, I. N. Aplicação da Lei dos Dígitos Significativos a Dados de Natureza Eletroquímica. In: Anais do XXIV Encontro Nacional de Modelagem Computacional e o XII Encontro de Ciência e Tecnologia de Materiais, 2021. Disponível em:<https://www.even3.com.br/anais/xxivenmc_xiiectm/404137-aplicacao-da-lei-dos-digitos-significativos-a-dados--de-natureza-eletroquimica/>. Acesso em: 01/08/2022.

[5] OLIVEIRA, C. A. L. S; MOURA NETO, F. D.; BASTOS, I. N. Mapas de difusão na busca de outliers de curvas de polarização. VETOR - Revista De Ciências Exatas e Engenharias, 32(1), 2–12. Disponível em: https://doi.org/10.14295/vetor.v32i1.13500. Acesso em: 01/08/2022.