Biomatemática ***34*** *(2024), 45–56* ***ISSN 1679-365X***
Uma publicação do grupo de Biomatemática IMECC – Unicamp

Utilizando agrupamento automático na identificação de campos vetoriais em sistemas p-fuzzy

Gislaine O. Queiroz[1], João B. Florindo[2], Estevão Esmi[3]
DMA, IMECC-Unicamp – 13.083-859, Campinas/SP.

Resumo. Este trabalho propõe uma alternativa para a realização de modelagem de sistemas dinâmicos, utilizando um sistema de base de regras *fuzzy* (SBRF), cujas regras são obtidas com o auxílio de um método de aprendizado de máquinas para clusterização, mais precisamente o método *K-means*.

Palavras-chave*: Sistema de base de regras fuzzy; sistema dinâmico; aprendizado de máquinas.*

1. Introdução

Propor uma modelagem para um campo vetorial de um sistema dinâmico não costuma ser uma tarefa fácil, pois há diversos obstáculos relacionados a incertezas intrínsecas do objeto de estudo. Deste modo, os Sistemas de Bases de Regras Fuzzy (SBRF) podem auxiliar na modelagem de campos vetoriais, por se tratar de uma ferramenta capaz de lidar com dados incertos e/ou parciais das dinâmicas estudadas. Esse tipo de abordagem de aproximar um campo vetorial parcialmente conhecido por um SBRF é denominada de sistemas parcialmente fuzzy ou, simplesmente, sistemas p-fuzzy (Barros et al., 2016). Uma outra vantagem de utilizar SBRF é que as regras dos sistemas podem ser interpretadas através de uma linguagem mais natural o que facilita a compreensão e a interdisciplinaridade entre especialista de áreas diversas.

[1]g155579@unicamp.br
[2]florindo@unicamp.br
[3]eesmi@unicamp.br

Além disso, quando dispomos de dados, a elaboração do sistema de regras pode ser assistida pelo uso de técnicas adequadas para tratar estes dados, tal como o emprego dos algoritmos de aprendizado de máquinas.

Neste contexto, o presente estudo propõe automatizar a geração de regras, de modo que os conjuntos fuzzy (Zadeh, 1965) sejam devidamente delimitados com o auxílio do método de aprendizado de máquinas *K-means*, cabendo ao especialista analisar esses conjuntos fuzzy e lhes atribuir significado.

Assim, em posse de um conjunto de dados referentes a um sistema dinâmico, os mesmos são organizados visando obter a variável de estado de interesse (y) e a sua derivada temporal $\left(\frac{dy}{dt}\right)$, resultando em um gráfico do tipo $y \times \frac{dy}{dt}$, cujos pontos são entradas do método de *K-means*.

Neste trabalho focaremos no SBRF do tipo Takagi-Sugeno-Kang (TSK), cujos antecedentes são conjuntos fuzzy e os consequentes são funções afins. Dado os agrupamentos produzidos pelo método *K-means*, determinamos os conjuntos fuzzy antecedentes utilizando a projeção dos agrupamentos obtidos no eixo horizontal do gráfico e, com o auxílio do método dos quadrados mínimos, os consequentes foram obtidos por retas que melhor se ajustam aos dados de cada grupo. Relacionando os conjuntos antecedentes e consequentes por regras *fuzzy* conjuntivas, obtem-se um SBRF do tipo TSK que estima o campo associado a dinâmica de y, isto é, que estima $\frac{dy}{dt}$ dado o estado de y. Em seguida, através do método numérico de Euler, conseguimos plotar a solução do nosso sistema p-fuzzy. A fim de comparar a qualidade da nossa estratégia de gerar automaticamente a base de regras de um sistema p-fuzzy assistido por um método de aprendizagem de máquina, iremos comparar a solução do sistema p-fuzzy produzido com a solução original de um problema de valor inicial conhecido.

2. Estado da arte

Dentro da literatura existente relacionada aos assuntos abordados aqui, traremos alguns dentre os trabalhos mais interessantes e de maior relevância. Por exemplo, quanto a utilização de sistemas fuzzy para auxílio de tomada de decisões referente s COVID-19, em Arora et al. (2021) os sintomas de algumas características relevantes dos indivíduos foram utilizados como antecedentes fuzzy para um sistema de inferência fuzzy, a fim de avaliar a necessidade de tratamento médico e isolamento o mais cedo possível, apresentando 97,2% de

precisão.

Alguns trabalhos têm surgido na linha do fuzzy K-means, de modo a suavizar a interseção dos agrupamentos, permitindo que um elemento possa pertencer a mais de um grupo. Wang et al. (2023) utilizaram essa abordagem para realizar análise de comportamento de preservação de privacidade em redes inteligentes, tendo alcançado resultados semelhantes aos do método tradicional, centrado em dados independentes, além de permitir a preservação da privacidade dos participantes.

Dentre os trabalhos mais recentes que estão relacionados com a utilização de SBRF e com métodos de aprendizado de máquinas, podemos destacar duas abordagens muito interessantes.

Em Sharaff et al. (2023) temos a utilização da ferramenta VADER em conjunto com sistema não supervisionado baseado em regras fuzzy, para fins de classificação das avaliações. A ferramenta VADER (*Valence Aware Dictionary and sEntiment Reasoner*), consiste em uma biblioteca de processamento de linguagem natural, utilizada para fazer a análise de textos escritos.

Assim, tal ferramenta foi utilizada para realizar a classificação das avaliações realizadas na *Amazon Fine foods*, atribuindo a cada palavra uma categorização positiva e uma negativa, podendo estas serem alta, média ou baixa. Essas duas são relacionadas através de um sistema experimental não supervisionado baseado em regras fuzzy, que consiste em nove regras para classificar as avaliações em positivas ou negativas.

Alcañiz et al. (2023) faz um comparativo da utilização de algoritmos de aprendizado de máquinas em conjunto com outras metodologias, para realizar a predição de energia fotovoltaica. O estudo avalia publicações realizadas entre 2000 e meados de 2022 e dentre essas, a abordagem fuzzy apareceu apenas duas vezes, de modo que a única vez na qual foi empregada com o método de *K-means*, foi utilizada para incluir uma interseção gradativa entre dois clusteres.

Embora seja um recorte dos trabalhos que utilizam aprendizado de máquinas para predição, esse trabalho nos dá uma dimensão de como estudos envolvendo lógica fuzzy e aprendizado de máquinas são escassos, além de não usufluir da interpretabilidade fornecida por métodos não supervisionados para fazer análises fuzzy.

No geral, podemos notar que há uma lacuna em relação a trabalhos que ofereçam uma alterativa para automatizar a geração de regras, bem como de trabalhos que relacionam sistemas p-fuzzy e algoritmos de aprendizado de má-

quinas. Neste sentido, nosso trabalho visa preencher esses hiatos e se apresenta com alternativa para geração automática de regras fuzzy, para descrição de campos vetoriais de sistemas dinâmicos.

3. Sistemas p-fuzzy

Utilizaremos um Sistema de Base de Regras *Fuzzy* (SBRF), que consite em um mapeamento do tipo $\Phi(\cdot) = D(I(F(\cdot)))$, em que F é o módulo de fuzificação, I é o módulo de inferência e D o módulo de defuzificação (Pedrycz e Gomide, 2007). No módulo de fuzzificação, contaremos com o método de inclusão canônica, que consiste em uma função que associa cada vetor $x \in \mathbb{R}^n$ com o subconjunto *fuzzy* $\{x\} \in \mathcal{F}(\mathbb{R}^n)$.

Para o módulo de inferência, estamos tomando o método (TSK) de Takagi-Sugeno-Kang (Takagi e Sugeno, 1985), que obedecerá ao formato (3.1)

$$y = \frac{\sum_{j=1}^{K} w_j g_j(x_1, ..., x_n)}{\sum_{j=1}^{K} w_j}, \ \forall\ j = 1, ..., K, \tag{3.1}$$

de modo que w_j indica o grau de pertinênica de $(x_1, ..., x_n)$ no j-ésimo antecedente, g_j a j-ésima função abordada que, neste caso, são retas, com $j = 1, ..., K$, sendo K o número de regras do SBRF.

Já para a defuzificação, tomamos o método do centro de massa $D(B)$ (Castro e Delgado, 1996). Mais precisamente, seja B um subconjunto *fuzzy*, no caso contínuo, o centro de massa $D(B)$ é dado por

$$D(B) = \frac{\int_X xB(x)dx}{\int_X B(x)dx}.$$

Cabe ainda ressaltar que utilizamos regras conjuntivas R_i, com $i = 1, ..., K \in \mathbb{N}$ (Zadeh, 1994), que obedecem a estrutura

$$R_i : \textbf{\textit{If}}\ x\ is\ A_i\ \textbf{\textit{then}}\ y = g_i(x), \qquad i = 1, ..., K \in \mathbb{N}$$

E por fim, para a obtenção da solução do sistema *p-fuzzy*, utilizaremos o método numérico de Euler (Barros et al., 2016).

$$x_{n+1} = x_n + \tilde{f}(t_n, x_n)h,$$

em que h é o tamanho do passo tomado e $\tilde{f}(t_n, x_n)$ é a saída do SBRF para a entrada (t_n, x_n).

4. Aprendizado de máquinas

O aprendizado de máquinas (*machine learning*) consiste em uma série de tarefas que são "ensinadas" para a máquina sem que esta tenha sido explicitamente programada para realizadar aquela tarefa (Bishop e Nasrabadi, 2006). Na maioria das vezes, tal processo de aprendizado se dá por exemplos.

Existem três grandes grupos de algoritmos de aprendizado de máquinas: o aprendizado supervisionado, o não supervisionado e o por reforço (Jane e Ganesh, 2019). Aqui será utilizado um método do tipo não supervisionado. Nesse tipo de abordagem, os dados processados não possuem qualquer tipo de rotulação prévia, buscando por padrões ocultos que serão úteis para auxiliar na resolução de problemas.

Essa capacidade de encontrar padrões nos fornece interpretabilidade, o que tem grande importância na presente proposta, pois permite ao especialista dar significado aos grupos gerados, por isso o interesse nesse tipo de método.

Sendo mais específico, utiliza-se o método de *K-means*, que requer como entrada apenas o número de agrupamentos desejados (K) e o conjunto de dados que se quer analisar. Então, ele realiza a separação dos elementos em K grupos disjuntos, indicando os elementos mais próximos por rótulos em comum (aqui, optamos por utilizar cores), além de indicar seus respectivos centroides, de modo que os centroides são os pontos que minimizam a distância dos elementos do grupo em relação a ele, e maximiza a distância em relação aos pontos não pertencentes ao grupo em questão.

Na figura 1 temos um exemplo da aplicação do método, em que à esquerda temos uma dispersão de pontos. Ao utilizar esses elementos como dados de entrada, junto com o número de agrupamentos desejados (tomamos $K = 4$), obtemos a imagem à direita, de modo que os elementos que pertencem ao mesmo grupo estão assinalados em cores iguais, além da indicação do centroide de cada grupo em cinza.

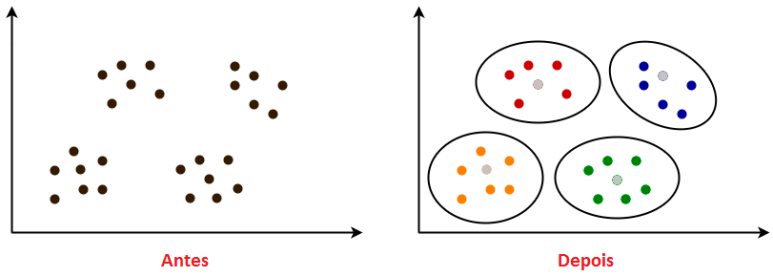


Figura 1: Exemplo de aplicação do método de *K-means*.

Note que o algoritmo não realiza qualquer mudança em relação aos elementos, apenas os identifica e mostra quais pertencem a qual grupo os sinalizando com o mesmo rótulo. Além disso, algumas configurações de dispersão podem não ser triviais e mesmo assim o método sempre converge (Bishop e Nasrabadi, 2006).

5. Metodologia

Neste estudo, estamos interessados em partir de um campo vetorial cujo formato seja semelhante ao da curva de crescimento populacional de Verhust, uma vez que esta é bastante conhecida na literatura, proporcionando meios de verificar e validar os resultados obtidos. Assim, precisamos organizar os dados para que a realização de comparações seja possível.

Logo, podemos partir de dados referentes a um sistema dinâmico (y), que possuam uma distribuição semelhante à curva de Verhult, ou podemos iniciar tomando um campo referente à média móvel de um evento, pois o acumulado desses fornecerá um gráfico análogo ao que buscamos.

Em seguida, definido o conjunto de dados (y) a ser tratado, encontramos a sua derivada $\left(\frac{dy}{dt}\right)$ utilizando o coeficiente angular fornecido pelo método dos quadrados mínimos. Assim, podemos obter um gráfico que relaciona o sistema dinâmico com a sua derivada $\left(y \times \frac{dy}{dt}\right)$. Esse gráfico contém os dados que, juntamente com o número de agrupamentos desejados (K), servirão como entrada para o método de *K-means*.

Então, o método de *K-means* fornece K clusterizações e podemos realizar suas projeções no eixo horizontal de modo a obter os conjuntos antecedentes, cujo centroide indica o ponto de maior pertinência da regra e também o ponto

cujas regras vizinhas não têm pertinência, de modo que estamos utilizando conjuntos trapezoidais para os extremos e triangulares para os demais.

Já para definir os conjuntos consequentes, tomamos o ajuste de retas fornecido pelo método dos quadrados mínimos, em relação a cada grupo. Agora, as regras antecedentes e consequentes são relacionadas através de regras fuzzy do tipo conjuntivas, com a utilização da inferência de Takagi-Sugeno-Kang, e tomamos o método de Euler para nos auxiliar a observar a solução obtida pelo sistema p-fuzzy. Por fim, a medida de RMSE auxilia a comparar o resultado com a curva original (y), com o intuito de ter um indicativo de quão próximos ou distantes estamos do esperado.

6. Resultados

O emprego do nosso método versa sobre os dados de um campo que representa um sistema dinâmico referente ao número de óbitos decorrentes da COVID-19, no período que corresponde de 25 de fevereiro de 2020 até 20 de junho de 2022. Fonte: `https://covid.saude.gov.br/`.

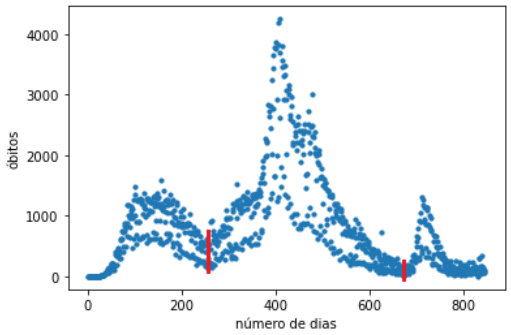
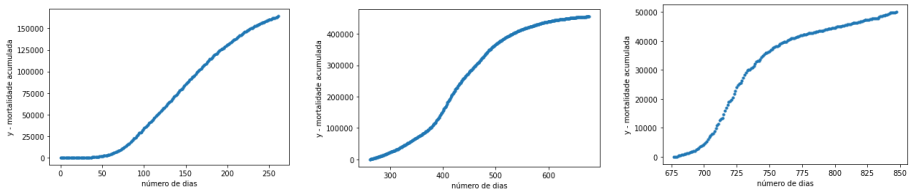


Figura 2: Gráfico de mortalidade populacional

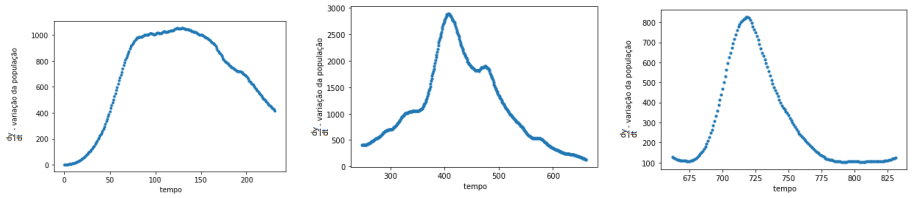
Os dados foram separados de acordo com as três curvas principais que correspondem cada uma a uma "onda" da doença, indicadas por barras verticais. A finalidade com essa divisão é fazer uma investigação mais profunda dos dados. Assim, uma vez que tais dados são análogos à média diária de óbitos, colocamo-nos em sequência, obtendo para cada onda estruturas semelhantes à curva de Verhulst, como podem ser verificadas na figura 3.



(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 3: Número de óbitos (y).

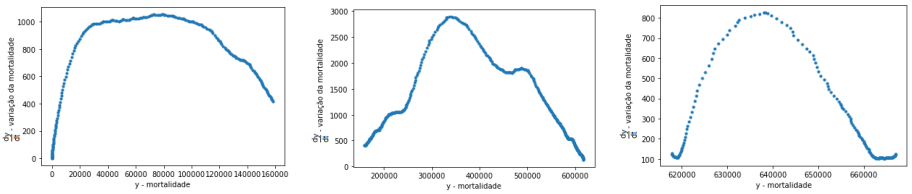
Assim, aplicamos o método dos quadrados mínimos aos elementos da figura 3 a fim de estimar suas respectivas derivadas temporais através dos coeficientes angulares fornecidos pelo método, resultando estão na figura 4.



(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 4: Derivadas temporais estimadas $\left(\frac{dy}{dt}\right)$.

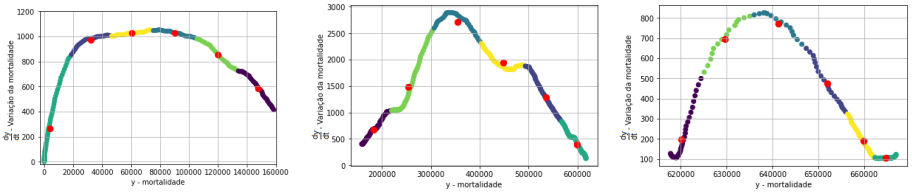
A seguir, relacionamos os gráficos da figura 3 com os da figura 4, obtendo então três gráficos do tipo $\left(y \times \frac{dy}{dt}\right)$, ilustrados na figura 5.



(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 5: Número de óbitos (y) pela derivada temporal $\left(\frac{dy}{dt}\right)$.

Os gráficos presentes na figura 5 são importantes pois são utilizados como dados de entrada para o método de *K-means*, juntamente com o número de agrupamentos desejados. Após a aplicação do algoritmo de agrupamento, recebemos um gráfico semelhante, com os grupos de interesse sinalizados por cores distintas. Esses grupos serão a base dos nossos conjuntos *fuzzy* antecedentes e consequentes. Logo, após utilizarmos o método de aprendizado de máquinas, chegamos à figura 6.



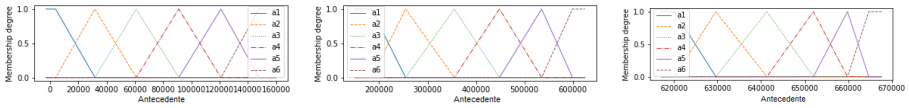
(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 6: Aplicação do método de *K-means*.

Na figura 6, tomamos o número $K = 6$ para a realização dos agrupamentos, de modo que esse número foi obtido empiricamente, através de testes que utilizavam diferentes números de grupos e diferentes configurações para os conjuntos *fuzzy* consequentes, e esse número de agrupamentos apresentou maior estabilidade e resultados mais próximos ao campo original. Esses experimentos são melhor detalhados em Queiroz (2023).

Ainda sobre a figura 6, fizemos a projeção dos conjuntos obtidos no eixo horizontal, de modo que o centroide de cada grupo indica o ponto de pertencimento máximo da regra, e o limite das mesmas será dado pelo tamanho da projeção de cada grupo, compondo assim as regras antecedentes, tal como na figura 7.

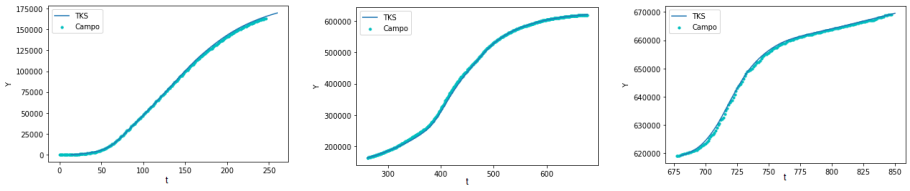
Já para encontrar os consequentes, utilizamos o método de Takagi-Sugeno, cujas regras são dadas pelos ajustes de retas que passam por cada grupo, de modo que cada grupo é descrito por uma reta diferente. Para esse ajuste de curvas linear, utilizamos o método dos quadrados mínimos.



(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 7: Conjuntos antecedentes.

Podemos então relacionar os conjuntos antecedentes e consequentes, obendo então um SBRF, com a inferência de Takagi-Sugeno e o médodo de defuzzificação do centroide. Com o auxílio do método numérico de Euler, obtemos a saída do SBRF, correspondendo a três curvas de Verhulst ajustadas. Os resultados podem ser conferidos na figura 8.



(a) Primeira onda (b) Segunda onda (c) Terceira onda

Figura 8: Soluções obtidas.

Para comparar os resultados na figura 8 com os originais que estão descritos na figura 3, foi utilizada a Raíz do Erro Quadrático Médio (RMSE, acrônimo do termo em inglês *Root Mean Squared Error*), que nos indica a proximidade entre ambas, sendo obtida pela fórmula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}},$$

em que x_i corresponde aos valores reais e y_i é referente aos valores obtidos através do nosso método, considerando série com n elementos. Os RMSEs resultantes das análises da primeira, segunda e terceira ondas foram $RMSE = 11.261,39$, $RMSE = 4.695,17$ e $RMSE = 1.601,82$ respectivamente, que são valores pequenos levando em consideração que estamos trabalhando com dados na ordem de 10^6.

Além desse aspecto quantitativo, também podemos notar que as curvas

geradas preservam todas as nuances das curvas originais. Logo, podemos considerar que tratam-se de boas aproximações também qualitativamente falando.

7. Conclusões

Para quantificar os resultados obtidos e ilustrados na figura 8 tomamos o cálculo do RMSE. Uma vez que resultaram em valores baixos de RMSE, isso é um indicativo de que as curvas adquiridas através da metodologia proposta geraram curvas próximas às originais. Assim, temos que a utilização de um SBRF com inferência de Takagi-Sugeno cujas regras são obtidas com o auxílio do método de aprendizado de máquinas *K-means* mostrou-se coerente e uma alternativa viável para a realização de modelagem de campos vetoriais para sistemas dinâmicos.

Referências

Alcañiz, A., Grzebyk, D., Ziar, H., e Isabella, O. (2023). Trends and gaps in photovoltaic power forecasting with machine learning. *Energy Reports*, 9:447–471.

Arora, S., Vadhera, R., e Chugh, B. (2021). A decision-making system for corona prognosis using fuzzy inference system. *Journal of fuzzy extension and applications*, 2(4):344–354.

Barros, L. C., Bassanezi, R. C., e Lodwick, W. A. (2016). *First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics*, volume 347 of *Studies in Fuzziness and Soft Computing*. Springer.

Bishop, C. M. e Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Castro, J. L. e Delgado, M. (1996). Fuzzy systems with defuzzification are universal approximators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1):149–152.

Jane, J. B. e Ganesh, E. (2019). A review on big data with machine learning and fuzzy logic for better decision making. *International Journal of Scientific & Technology Research*, 8(10):1121–1125.

Pedrycz, W. e Gomide, F. (2007). *Fuzzy Systems Engineering: Towards Human-Centric Computing*. Wiley, IEEE Press, New York.

Queiroz, G. O. (2023). Utilização de aprendizado de máquinas na identificação de campos vetoriais em sistemas p-fuzzy. Dissertação de Mestrado, IMECC–Unicamp, Campinas/SP.

Sharaff, A., Rajput, N., e Papatla, S. R. (2023). Unsupervised sentiment analysis of amazon fine food reviews using fuzzy logic. In *International Conference on Computing, Communication and Learning*, páginas 126–137. Springer.

Takagi, T. e Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, SMC-15(1):116–132.

Wang, Y., Ma, J., Gao, N., Wen, Q., Sun, L., e Guo, H. (2023). Federated fuzzy k-means for privacy-preserving behavior analysis in smart grids. *Applied Energy*, 331:120396.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Zadeh, L. A. (1994). Soft computing and fuzzy logic. *IEEE software*, 11(6):48–56.