Mapas Auto-Organizáveis Aplicado à Análise de Poluentes Atmosféricos na Cidade de Salvador, Bahia

Emanoel L. R. Costa*, Taiane Braga†, Édler L. de Albuquerque‡ e Marcelo A. C. Fernandes*§
*Laboratório de Aprendizagem de Máquina e Instrumentação Inteligente, nPITI/IMD, UFRN, Natal, RN, Brasil.
†IFBA, Salvador, BA, Brasil.
‡Departamento de Processos Industriais e Engenharia Química (DEPEQ), IFBA, Salvador, BA, Brasil.
§Departamento de Engenharia da Computação e Automação, UFRN, Natal, RN, Basil.
Email: *lucas.rodrigues@dca.ufrn.br,†taianecbraga@gmail.com,‡edler@ifba.edu.br,§mfernandes@dca.ufrn.br

***Resumo*—A poluição atmosférica é um problema que está cada vez mais presente em nossa sociedade devido ao crescente desenvolvimento dos países. No estudo de poluentes atmosféricos, métodos de estatística multivariada são comumente utilizados, porém a aprendizagem de máquina tem se mostrado uma ótima alternativa, dispondo de técnicas capazes de lidar com problemas de grande complexidade, como é o caso da poluição. Neste trabalho foi utilizada uma técnica baseada em uma rede neural artificial não supervisionada do tipo Mapas Auto-Organizáveis, com o objetivo de realizar uma nova abordagem para a análise de dados de poluentes atmosféricos. A análise foi realizada na cidade de Salvador – Bahia em uma única estação de monitoramento da qualidade do ar localizada no bairro Itaigara, com base em dados medidos durante os anos de 2011 a 2016 e disponibilizados pelo Governo do Estado da Bahia por meio da empresa CETREL S. A.. A partir dos resultados obtidos pela aplicação dos Mapas Auto-Organizáveis foi possível realizar um estudo exploratório sobre os dados de poluição, evidenciado correlações entre poluentes e dados meteorológicos, agrupamento de dados, possíveis fontes de emissão e conclusões sobre o problema.**

***Index Terms*—Aprendizagem de Máquina, Poluentes Atmosféricos, Mapas Auto-Organizáveis, SOM, Salvador-BA**

I. Introdução

A poluição do ar é um dos grandes desafios existentes na sociedade moderna. Nos últimos anos, a poluição causada por fontes de emissão industrial, veicular e de produtos químicos tóxicos vem aumentando significativamente, e esse aumento pode ser visto sobretudo em países de baixa e média renda, os quais se enquadram nos chamados países em desenvolvimento [1]. Apesar de possuírem essa tendência crescente na magnitude de poluição, agendas destinadas a conscientização ou programas de controle à poluição ainda recebem pouca atenção e recursos, seja do governo, de agências internacionais ou doadores filantrópicos [1].

Efetuar uma gestão e um controle efetivo da poluição do ar requer um grande conhecimento acerca dos custos, assim como dos benefícios dessa regulamentação. Os principais esforços voltados para a mensuração de poluentes têm por objetivo evitar possíveis malefícios à saúde, como doenças respiratórias ou cardiovasculares que podem resultar em hospitalizações e até morte, atingindo normalmente segmentos vulneráveis da população [2].

Os componentes que contribuem para a poluição do ar são misturas complexas, compostas tanto por partículas sólidas, como por poluentes gasosos. Dentre esses, existem os poluentes atmosféricos prioritários, os quais são geralmente objeto de regulamentação por força da lei. Os principais poluentes considerados prioritários no Brasil incluem substâncias que podem ser lançadas diretamente na atmosfera (poluentes primários) e substâncias formadas a partir dos primários por meio de reações fotoquímicas na troposfera (poluentes secundários) [3]. Os gases poluentes incluem gases como o dióxido de enxofre (SO_2), dióxido de nitrogênio (NO_2), monóxido de carbono (CO), e compostos orgânicos voláteis (COVs), bem como materiais sólidos ou líquidos, suspensos na atmosfera devido ao seu pequeno tamanho, chamados material particulado (MP). Adicionalmente, o ozônio (O_3), um dos principais poluentes fotoquímicos formado na atmosfera pela reação de óxidos de nitrogênio (NO_x) e hidrocarbonetos como os COVs na presença de luz solar, similarmente ao sulfato particulado e aos aerossóis de nitrato, criados a partir do SO_2 e NO_x [3].

Condições meteorológicas como a temperatura, umidade relativa, pressão atmosférica, direção e velocidade do vento, como também condições do terreno podem afetar na dispersão de poluentes atmosféricos [4]. O que dificulta a análise e identificação de poluentes e fontes principais em áreas de grande escala, que remete também a um problema de posicionamento das estações de monitoramento para coleta de dados.

Diversas são as fontes de emissão de poluentes do ar. De forma geral, uma única fonte é capaz de emitir vários poluentes, como é o caso dos veículos automotores e, dependendo da composição dos combustíveis fósseis, diferentes poluentes podem ser emitidos pela sua combustão e evaporação, bem como pelo desgaste de pneus e estradas onde os veículos circulam. A emissão veicular, devido ao seu crescente aumento no número de veículos privados, tem se tornado uma fonte dominante na emissão de CO, CO_2, COVs, NO_x e MP, ao passo que processos industriais normalmente incluem poluentes como CO, MP, NO_x e SO_2 [4]–[6].

Assim, a observação da concentração de poluentes no meio ambiente em determinados pontos de monitoramento torna-se uma tarefa essencial e necessária. Identificar as principais componentes presentes possibilita a obtenção de conhecimento sobre a situação atual da poluição do ar, variações, correlações, possíveis fontes de emissão e conduz ao desenvolvimento de políticas públicas de conscientização e redução de poluentes. Nesse sentindo, muitas pesquisas se propuseram à análise de dados ambientais utilizando principalmente técnicas de estatística multivariada [4], [7], [8].

Métodos da estatística multivariada como a análise por correlação [9], [10], análise de *clusters* [11]–[13] e a análise de componentes principais [7], [14], [15], são comumente aplicados em diversos estudos na identificação de relações entre poluentes ou outras variáveis que podem influenciar na qualidade do ar. Trabalhar com grandes bases de dados que carregam múltiplas informações sobre o espaço observado, como é o caso da poluição do ar, requer a utilização de técnicas que proporcionem a obtenção, extração e identificação de características inerentes aos dados analisados.

Nesse contexto, a aprendizagem de máquina tem se mostrado uma ótima alternativa aos métodos comuns utilizados [16]. Um algoritmo bastante conhecido e que faz parte do grupo dos algoritmos de aprendizagem não-supervisionada é o de Mapas Auto-Organizáveis (*Self-Organizing Maps* – SOM) [17]. Por ser um método de redução de dimensionalidade para análise de dados, o SOM pode apresentar semelhanças com as técnicas multivariadas amplamente utilizadas, no entanto, se comparado com as mesmas, apresenta vantagens como a não necessidade de fazer suposições sobre a distribuição das variáveis, é capaz lidar com problemas não-lineares de grande complexidade e dimensão, e é eficaz na utilização de dados ruidosos [18].

Com relação às aplicações do SOM sobre dados de poluentes atmosféricos, as investigações se voltam para diversas aplicações. A qualidade do ar é sempre um fator importante e sua avaliação por meio do SOM permite uma interpretação mais fácil dos resultado e classificação, conforme demonstrado por [19]. Em [20] os autores utilizaram do SOM para identificar os níveis de poluição presente na exposição durante a mineração de terra e fundição. O estudo realizado em [21] utilizou o SOM para evidenciar como diferentes tipos de circulação do ar podem causar efeitos na qualidade do ar de certas regiões, alterando a concentração de poluentes presentes na atmosfera. A identificação de locais adequados para o posicionamento de estações de monitoramento também é uma etapa essencial e que deve ser avaliada, como mostrado em [22]. Os trabalhos [23] e [24] utilizaram do SOM parar obter características relacionadas ao material particulado presente na atmosfera, avaliando fatores presentes nas concentração encontradas tanto na exposição interna como externa, relacionando-os com as atividades humanas. Para lidar com a avaliação de zonas de poluição por ozônio, conforme demonstrado por [25], o SOM também pode funcionar como identificador de poluição, definindo limites para classificar regiões com baixa ou alta concentração de determinado poluente como o ozônio.

Com o objetivo de realizar uma análise de dados de poluentes atmosféricos, neste trabalho é utilizado o método baseado no modelo de redes neurais artificiais, SOM. Sendo o mesmo uma técnica de projeção não linear de um espaço de alta dimensão para baixa dimensão, que possibilita uma análise acerca de agrupamentos, correlação, interpretação visual e classificação de dados. O SOM foi aplicado a uma base de dados de poluentes medidos por uma estação de monitoramento do ar, dispondo de amostras de SO_2, CO, O_3, MP, NO, NO_2, velocidade do vento, temperatura, umidade relativa e desvio-padrão da direção do vento, localizada na cidade de Salvador, Bahia. O intuito é realizar uma análise acerca destes poluentes, buscando identificar padrões, correlações e características entre os dados estudados.

II. Materiais e Métodos

A. Área de Estudo

Salvador, capital do estado da Bahia, possui uma área territorial de 693,453 km^2 e conta com uma população de 2.675.656 de pessoas. Situado na região nordeste do Brasil, 12° 58' 16" de latitude sul e 38° 30' 39" de longitude oeste, é uma cidade de núcleo urbano e topografia acidentada formada por diversas colunas e vales, com um clima tropical chuvoso, sem estação seca e temperatura média anual por volta de 25° C. Este trabalho utilizou uma base de dados disponibilizada pelo Governo do Estado da Bahia por meio da empresa CETREL S. A., e que foi proveniente da rede de monitoramento da qualidade do ar implantada na cidade de Salvador. Assim sendo, para a análise e estudo aqui desenvolvido, foi escolhida somente uma estação específica, a chamada Estação Itaigara. A Fig. 1 ilustra como estavam distribuídas as demais estações (E1 - E7) e destaca a estação de estudo.

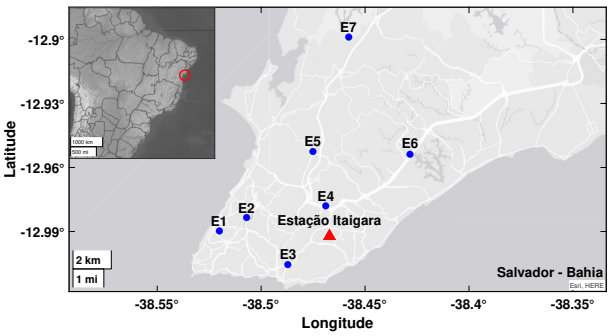

Figura 1. Localização das estações de monitoramento em Salvador-BA.

A região de Itaigara, está localizada na parte sudeste da cidade, sendo caracterizada por ser uma área comercial e residencial com um intenso fluxo de veículos, dispondo de prédios como shoppings, supermercados, condomínios, hospitais, universidade e fazendo ligação com importantes saídas como o aeroporto, a BR-324 e a área da rodoviária. A estação de monitoramento utilizada foi, durante seu funcionamento, alocada em um ponto estratégico entre duas avenidas, Av. Antônio Carlos Magalhães e Av. Juracy Magalhães Jr., devido ao constante tráfego e ligações que as mesmas possuem.

B. Base de Dados

A CETREL S. A. operou as estações de monitoramento durante o período que compreendeu os anos de 2011 a 2016. A base de dados conta com registros de 12 variáveis referente às médias horárias de parâmetros meteorológicos e concentrações de poluentes. Em relação aos parâmetros meteorológicos a base dispõe dos parâmetros velocidade do vento (VV), temperatura ambiente (TEMP), umidade relativa do ar (UR), desvio-padrão da direção do vento (STWD), chuva e direção do vento. Já para os poluentes tem-se o SO_2, CO, O_3, material particulado cujo diâmetro aerodinâmico é menor que $10\mu m$ (MP_{10}) e os óxidos de nitrogênio NO_2 e NO. Devido a pequena quantidade de dados de chuva e direção do vento, os mesmos foram excluídos do estudo realizado, perfazendo assim um total de 10 variáveis analisadas. A estação estudada operou durante outubro de 2013 a abril de 2016, registrando originalmente um total de 22.203 amostras dessas 10 variáveis.

Para evitar possíveis erros de medições e obter uma melhor qualidade na análise foi realizado um pré-processamento dos dados, dividido em três etapas. A primeira delas trata-se da remoção de linhas nulas, seguida logo em sequência pela segunda etapa, na qual erros de medições provenientes dos equipamentos utilizados foram identificados e removidos, devido a sua nomenclatura específica na base de dados. Restando ao final da primeira e segunda etapa de 17.078 amostras.

A terceira e última etapa de pré-processamento foi caracterizada pela remoção de *outliers*. Nesta etapa foi realizada uma investigação sobre dispersão e simetria dos dados e foi empregada a medida de separatriz por quartis [26], a qual divide o conjunto de dados em três partes definidas como quartis Q_1, Q_2 e Q_3. Como critério para identificação de *outliers* das variáveis foram considerados somente os *outliers* extremos, ou seja, que possuem valor maior do que $Q_3 + 3 \times AIQ$ e menor do que $Q_1 - 3 \times AIQ$, onde AIQ é a amplitude interquartil [26], sendo estes removidos da base dados. Os *outliers* moderados, valor maior do que $Q_3 + 1.5 \times AIQ$ e menor do que $Q_1 - 1.5 \times AIQ$, foram mantidos para evitar uma redução grande do conjunto de dados.

Ao fim do pré-processamento foi obtido um total de $P = 15.535$ amostras, as quais foram de fato utilizadas para o estudo. A Tabela I apresenta a estatística descritiva de todos os poluentes e dados atmosféricos medidos pela estação Itaigara em Salvador sobre as $P = 15.535$ amostras, com concentração dos poluentes em partes por bilhão (ppb).

O (SO_2), resultante da queima de combustíveis com enxofre apresenta a menor concentração, tendo máxima de 1,6 ppb e média de 0,25 ppb, ao passo que as maiores concentrações foram de CO com máxima de 1.210 ppb e média de 226,48 ppb, que resulta da queima de combustíveis com origem orgânica. Os óxidos de nitrogênio (NO e NO_2) formados durante processos de combustão e reações químicas atmosféricas apresentaram, respectivamente, máximas de 70,70 ppb e 31,10 ppb e médias de 15,50 ppb e 8,21 ppb. O O_3, com máxima de 27,50 ppb e média 8,47 ppb, indica a presença de oxidantes fotoquímico e o MP_{10}, material sólido que se mantém suspenso na atmosfera, com concentração máxima de 67,40 $\mu g/m^3$ e média 16,16 $\mu g/m^3$.

As maiores variações nas concentrações dos poluentes estão no SO_2, CO e NO. Isso pode se dar devido a localização em que se encontra a estação de monitoramento, em meio duas avenidas de trânsito de veículos, já que esses poluentes são principalmente emitidos pela queima de combustíveis. O trânsito pode ser intenso em determinadas horas do dia e mais calmo em outras, no período que compreende as 24 horas diárias de coleta de dados.

Um fato importante é que os resultados do SOM podem ser afetados por determinadas variáveis que possuem uma escala maior do que as demais, sendo assim, métodos de reescalonamento de dados como a normalização *Min-Max*, que reescala os valores para o intervalo entre 0 e 1, a normalização z-score, tornando os dados com média nula e variância unitária e a transformação *logarítmica* reduzindo a escala dos dados, foram utilizadas em fase preliminar e analisadas quanto ao desempenho proporcionado, discutidos mais adiante.

C. Mapas Auto-Organizáveis (Self-Organizing Maps - SOM)

Self-Organizing Map, Mapas Auto-organizáveis, Mapas de Kohonen ou Redes de Kohonen, são nomes comumente utilizados para apresentar um mesmo método, mais popularmente conhecido somente por SOM, proposto por [17]. O SOM é um modelo de rede neural com característica de projeção não linear de um espaço de entrada de alta dimensão para um espaço de baixa dimensão, com uma superfície de saída composta por unidades ordenadas chamadas de neurônios. Sendo um método bastante aplicado em problemas que envolvem redução não linear de dimensionalidade e agrupamento de dados [27].

A rede SOM consiste de um conjunto de M neurônios dispostos em um arranjo L-dimensional, chamado de mapa. Todos os M neurônios estão conectados a uma mesma entrada que pode ser caracterizada por um vetor $\mathbf{x}$ N-dimensional expresso como $\mathbf{x} = [x_1, x_2, \ldots, x_N]$. Para um arranjo bidimensional ($L = 2$), a topologia do mapa que representa a rede SOM pode ter formatos do tipo retangular, hexagonal, quadrado e outros [17]. Cada i-ésimo neurônio do mapa possui um vetor de pesos de mesma dimensionalidade do vetor entrada, $\mathbf{x}$, expresso como $\mathbf{w}_i = [w_{i1}, w_{i2}, \ldots, w_{iN}]$. No caso arranjo bidimensional ($L = 2$), a topologia de um mapa com M neurônios pode ser expressa como $(M_h \times M_v)$ onde M_h é o número de neurônios na horizontal e M_l é o número de neurônios na vertical ($M = M_h \times M_v$).

O algoritmo do SOM consiste em um processo iterativo que organiza os neurônios de acordo com suas similaridades, realizando um mapeamento não linear dos dados de entrada e mantendo sua forma topológica. Dessa forma, neurônios adjacentes formam padrões e cada amostra é associada a um deles de acordo com sua similaridade [17]. O algoritmo consiste nos seguintes passos:

1) Inicializar os vetores de pesos dos M neurônios com valores aleatórios.
2) Calcular a distância de cada entrada n-ésima amostra de entrada $\mathbf{x}(n)$ para todos os M neurônios.

Tabela I
ESTATÍSTICA DESCRITIVA DOS POLUENTES E DADOS ATMOSFÉRICOS DA ESTAÇÃO ITAIGARA ($P = 15.535$ AMOSTRAS)

Variáveis	Unidade	Máximo	Mediana	Média	Desvio padrão	Coeficiente de Variação
SO_2	ppb	1,60	0,10	0,2502	0,33	131,89%
CO	ppb	1.210,00	190,00	226,48	207,26	91,51%
O_3	ppb	27,50	7,90	8,47	4,32	51,00%
MP	$\mu g/m^3$	67,40	13,60	16,16	10,98	67,94%
NO	ppb	70,70	11,40	15,50	13,45	86,77%
NO_2	ppb	31,10	7,30	8,21	5,15	62,72%
Velocidade do vento	m/s	10,20	2,70	2,76	1,58	57,24%
Temperatura	°C	33,40	25,00	25,04	2,27	9,06%
Umidade	%	93,00	71,00	71,43	9,08	12,71%
STWD	°	51,30	22,80	24,32	8,13	33,42%

3) Encontrar o j-ésimo neurônio vencedor a partir métrica expressa como

$$j = \arg\min_i ||\mathbf{x}(n) - \mathbf{w}_i(n)||, \; i = 1, 2, \ldots, M. \quad (1)$$

O neurônio vencedor é chamado de melhor unidade correspondente (*Best Matching Unit* - BMU).

4) Atualizar os pesos do BMU e de seus neurônios vizinhos de acordo com a equação expressa como

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \eta(n) h_{i,j}(n)(\mathbf{x}(n) - \mathbf{w}_i(n)) \quad (2)$$

onde $\eta(n)$ é a taxa de aprendizagem (variando entre 0 e 1) na n-ésima iteração e $h_{i,j}(n)$ é a função de vizinhança do i-ésimo neurônio para o BMU que é expressa como

$$h_{i,j}(n) = \exp\left(-\frac{d_{i,j}^2}{2\sigma^2(n)}\right) \quad (3)$$

onde $d_{i,j}^2$ a distância do i-ésimo neurônio para o BMU (j-ésimo neurônio) e $\sigma^2(n)$ indica o raio de influencia do BMU com seus neurônios vizinhos na n-ésima iteração.

5) Repetir os passos 2, 3 e 4 até atingir o número de máximo de iterações, chamado aqui de T.

Para um problema com P amostras de entrada, o número de iterações deve ser suficiente para aplicar várias vezes todas P amostras, ou seja, $T = b \times P$ onde b é número de vezes que todo conjunto de P amostras é apresentado para a SOM.

Conforme n aumenta, o raio da função de vizinhança, $\sigma^2(n)$, diminui gradativamente, assim somente aqueles neurônios que estão mais próximos do BMU serão atualizados, tornando a conexão entre eles cada vez mais forte e compartilhando de similaridades. Após treinada a rede, cada n-ésima entrada $\mathbf{x}(n)$ é associada a um determinado BMU na camada de saída, entradas que compartilham de padrões similares serão associados a um mesmo BMU ou aos seus vizinhos, o que pode ser entendido como agrupamento no SOM.

Nesse estudo, foram utilizados $P = 15.535$ amostras no qual cada i-ésima amostra é caracterizada por 6 variáveis de poluentes atmosféricos (SO_2, CO, O_3, MP_{10}, NO e NO_2) e 4 variáveis meteorológicos (VV, TEMP, UR e STWD), formando um espaço de dimensão $N = 10$. O intuito é realizar uma análise, encontrar características e discutir acerca dessa variáveis fazendo uso dos recursos de visualização de modelos, redução não-linear de dimensionalidade e agrupamento fornecidos pela rede SOM.

D. Determinação dos Parâmetros do Mapa

Na utilização do SOM uma etapa essencial é a definição do tamanho do mapa. É necessário determinar o número de neurônios que serão utilizados no treinamento de forma a evitar grandes ou pequenas quantidades de neurônios, as quais podem causar problemas como a não identificação de caraterísticas e *overfitting* [28]. Uma regra geral utilizada para determinar essa quantidade é a partir da equação heurística definida como

$$M \approx 5\sqrt{P} \quad (4)$$

onde P é o total de amostras [28]. Como este trabalho possui $P = 15.535$ amostras, tem-se $M = 623$.

Determinada a quantidade de neurônios, a melhor escolha para os valores de M_h e M_v relativos a topologia da SOM foi definida de acordo com medidas de qualidade usualmente utilizadas para a rede SOM, o Erro de Quantização (EQ) e o Erro Topográfico (ET) [29], [30]. Como apresentado na Tabela II, foram testados vários valores de M_h e M_v mantendo o valor $M = 623$. Para análise dos resultados, foram aplicadas aos dados 3 diferentes tipos de normalização: z-score, Min-Max e a Logarítmica.

Os testes foram realizados para uma topologia do tipo hexagonal e o algoritmo de treinamento foi aplicado em duas etapas, sendo $b = 500$ em ambas. Na primeira foi utilizado como valores iniciais $\eta(0) = 0.9$ e $\sigma^2(0) = \frac{M_h}{2}$, já na segunda etapa (refinamento) os valores foram fixados em $\eta = 0.05$ e $\sigma^2 = 1$. A Tabela II apresenta as medidas de qualidade obtidas para cada teste.

Considerando ambas as medidas conjuntas, os menores valores para EQ e ET foram obtidos fazendo uso da normalização *Min-Max* no mapa (25×25). Dessa forma, com base nos valores de qualidade obtidos e no número de neurônios mais próximo do calculado pela Eq. 4, foi aderido o mapa de tamanho (25×25) com $M = 625$ neurônios.

Tabela II
MEDIDAS DE QUALIDADE DO SOM

$(M_h \times M_v)$	M	z-score		Min-Max		Logarítmica	
		EQ	ET	EQ	ET	EQ	ET
(24×23)	552	1,4306	0,0584	0,2428	0,0591	0,7736	0,0510
(26×22)	572	1,4237	0,0603	0,2422	0,0566	0,7709	0,0485
(24×24)	576	1,4210	0,0618	0,2421	0,0557	0,7704	0,0503
(27×22)	594	1,4192	0,0548	0,2403	0,0589	0,7684	0,0547
(25×24)	600	1,4152	0,0593	0,2412	0,0565	0,7659	0,0477
(27×23)	621	1,4126	0,0574	0,2400	0,0585	0,7654	0,0444
(25×25)	625	1,4063	0,0573	**0,2399**	**0,0553**	0,7625	0,0458
(27×24)	648	1,4086	0,0572	0,2381	0,0561	0,7595	0,0472
(26×26)	676	1,3945	0,0640	0,2371	0,0556	0,7553	0,0525
(27×26)	702	1,3861	0,0578	0,2363	0,0559	0,7516	0,0538

III. RESULTADOS E DISCUSSÕES

A. U-matrix, Plano de Componentes e Similaridade entre as Variáveis

Definidos todos os requisitos necessários, o SOM foi aplicado à base de dados para processamento e obtenção da saída, a qual pode ser representada por dois tipos de mapas: matriz de distância unificada (*U-matrix*) e os planos de componentes, ambos ilustrados na Fig. 2. A *U-matrix* fornece uma visualização da distância relativa entre os neurônios no mapa, essa distância é evidenciada por meio de uma escala de cores e mostra a distância calculada entre os neurônios adjacentes [17]. Quanto mais a cor aproxima-se de um azul escuro na *U-matrix* mais próximos estão esses neurônios, ou seja, possuem uma maior similaridade, e quanto mais a cor aproxima-se de um vermelho escuro maior é distância entre esses neurônios, maior dissimilaridade. Assim, essa forma de representação permite considerar que neurônios com menores distâncias formam um *cluster*, enquanto neurônios com distâncias elevadas podem ser considerados como limites de um *cluster*.

Os planos de componentes mostram os valores dos vetores de pesos de cada neurônio por meio de um código de cores, onde as cores azul e vermelho correspondem a baixos e altos valores, respectivamente. Essa representação permite que a partir de uma comparação entre os padrões de cada plano possa ser feito um reconhecimento de dependências entre as variáveis. O gradiente de cor dos planos representa o nível das variáveis (componentes) para as amostras analisadas, a cada neurônio é atribuída uma cor de acordo com valor relativo da respectiva variável nesse neurônio, assim, pode-se dizer que duas ou mais variáveis são próximas baseado em uma comparação entre seus gradientes de cor. Um gradiente coerente indica uma correlação positiva, enquanto um gradiente inverso uma correlação negativa.

Analisando os planos de componentes da Fig. 2, dois padrões em especial são bastante perceptíveis, os planos de Umidade e Temperatura apresentam gradientes totalmente inversos, indicando uma correlação negativa entre essas variáveis, algo que já esperado dada as características das mesmas. Para os poluentes CO, NO, NO_2 os seus vetores de pesos apresentam altos valores que estão localizados na parte lateral esquerda do plano de componentes, indo do meio ao topo do mapa, porém com uma maior concentração de altos valores dos três planos na parte superior esquerda, evidenciando uma certa similaridade entre eles. Poluentes estes que são causados principalmente pela combustão, queima incompleta de combustíveis orgânicos, muito comum em cidades com grande circulação de veículos, principal emissor.

O O_3 que é formado por reações entre óxidos de nitrogênio e COVs apresenta um padrão diferente do NO_2, que tem uma contribuição importante na formação de oxidantes fotoquímicos como o O_3. Sua região de altos valores localizada no lado direito do plano exibi uma certa semelhança ao gradiente do plano de componente da Velocidade do vento. Essa tendência se deve ao fato do O_3 ser gerado por reações fotoquímicas e geralmente sua concentração é baixa próximo a vias de tráfego devido o consumo de O_3 pelo o NO. Assim, é mais provável que o O_3 registrado tenha vindo de outra região transportado pelo vento.

O MP apresentou um padrão diferente dos demais poluentes. Sua principal região de concentração de valores altos do vetor de pesos está na parte superior do plano, e sendo suas fontes de emissão diversas como veículos, queima de biomassa, indústrias, ressuspensão de poeira, entre outras, é difícil a identificação de um responsável principal. Apesar de que sua formação também pode ser realizada na atmosfera por meio de COVs, SO_2 e óxidos de nitrogênio.

O padrão mais distinto apresentado foi pelo SO_2, com valores altos e bem concentrados na parte inferior esquerda, não se assemelhando a nenhum outro plano de componentes. Um motivo para que isto tenha ocorrido é devido sua geração em ambiente urbano se dever a queima de óleo diesel em veículos pesados (caminhões, ônibus, micro-ônibus e parte dos comerciais leves)

A comparação de planos de componentes nos possibilita uma análise de como se relacionam as variáveis de forma simples e visual, a partir da identificação de padrões nos planos pode-se associar e assumir conclusões referentes ao nosso problema de acordo com o resultado obtido. Pelo SOM, sabe-se que padrões similares são dispostos em uma mesma região de vizinhança resultando em agrupamento na saída da rede, dessa forma uma investigação sobre o agrupamento das amostras fornece informações importante sobre os dados.

A *U-matrix* na Fig. 2 ilustra o quão próximo e o quão distante os neurônios estão, evidenciando assim os agrupamentos. Porém os limites dos *clusters* não estão claramente representados, dificultando a identificação. Uma alternativa para escolher a quantidade de *clusters* adequada é a utilização o índice de *Davies-Bouldin* que indicará o melhor número possível para representar o agrupamento das amostras.

B. Agrupamento das Amostras com o SOM

Com o objetivo de encontrar o número ótimo de *clusters* para os neurônios do mapa foi calculado o índice de *Davies-Bouldin* [31], no qual a melhor quantidade de *clusters* é aquela que apresenta o menor índice. Dessa forma, foi realizado um experimento variando o número de *clusters* no intervalo de 2

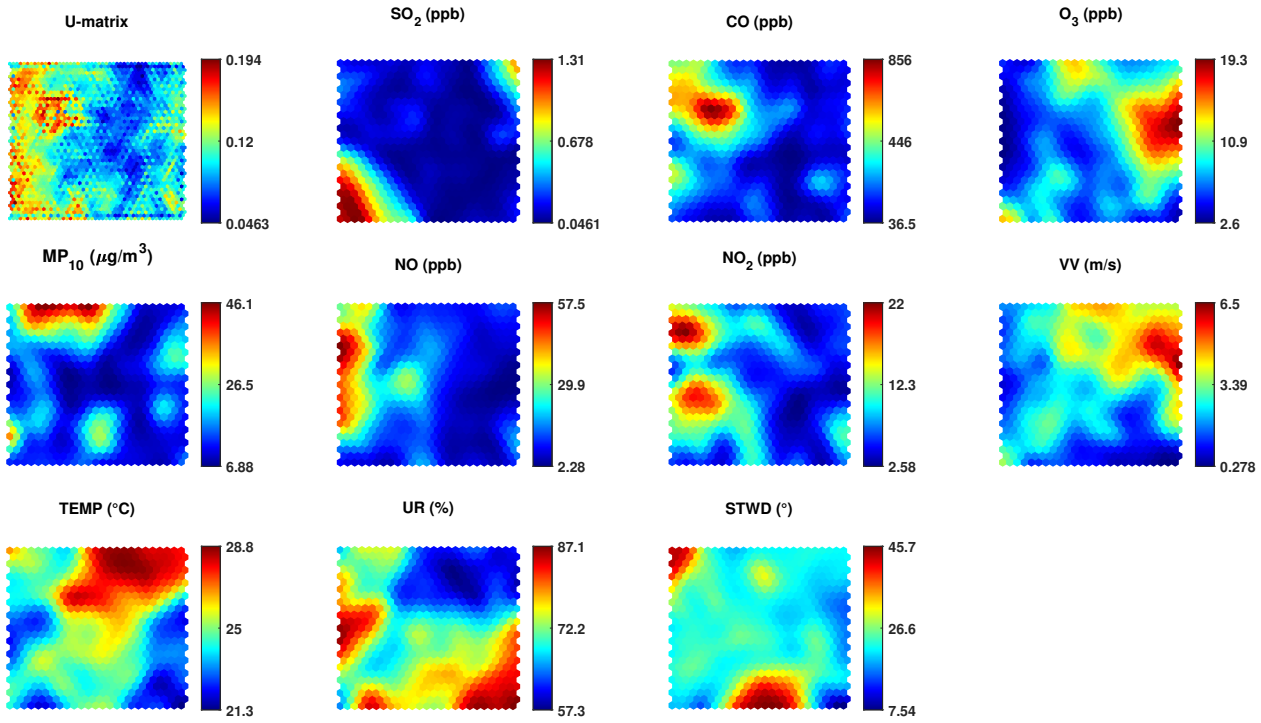

Figura 2. Matriz de distância unificada (*U-matrix*) e planos de componentes de todas as variáveis analisadas (SO_2, CO, O_3, MP_{10}, NO, NO_2, VV, TEMP, UR e STWD).

a 8 conforme apresentado na Fig. 3. Com base na curva obtida observa-se que o melhor resultado é alcançado para 4 *clusters*.

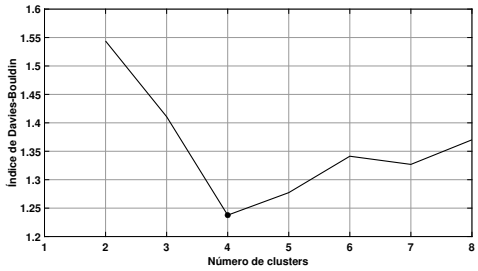

Figura 3. Valores do índice de *Davies-Bouldwin* obtidos para os diferentes números de *clusters* no intervalo de 2 a 8.

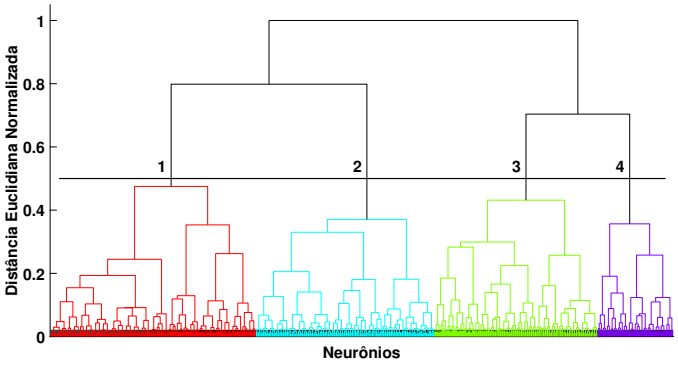

Figura 4. Análise hierárquica do agrupamento dos neurônios do mapa utilizando o método de ligação de *Ward* e distância Euclidiana.

A partir do número ótimo de *clusters* encontrado foi realizada uma análise hierárquica para definir os neurônios pertencente aos 4 *clusters*. Utilizando a distância euclidiana como medida de similaridade entre os neurônios e o critério de ligação de *Ward*, foi obtido o dendrograma ilustra na Fig. 4. No qual é definido um valor limite para classificar cada neurônio como pertencente a um determinado *cluster* (linha horizontal na Fig. 4).

Fundamentado na análise hierárquica os neurônios do SOM foram classificados e formaram os 4 *clusters* obtidos, conforme ilustra a Fig. 5. Assim sendo, cada *cluster* possui uma determinada quantidade de amostras atribuídas aos seus neurônios, as quais apresentam características sobre a distribuição dos poluentes e parâmetros meteorológicos. A Tabela III apresenta as médias das amostras das variáveis de acordo com o *cluster* as quais foram atribuídas.

De acordo com a Tabela III, as amostras pertencentes ao *cluster* 1 exibem, no geral, uma baixa concentração dos poluentes do ar, com exceção do O_3 e do MP. O O_3 possui a maior concentração média dentre todos os outros, assim como o MP é o segundo maior. A velocidade do vento é consideravelmente maior aqui, junto com a maior temperatura e menor umidade relativa. No total, cerca de 34% dos dados foram atribuídas ao *cluster*, compartilhando dessas características.

O *cluster* 2 na Tabela III apresenta as menores concentrações dos poluentes SO_2, CO, MP e NO, com valores intermediários de O_3 e NO_2. Possuindo a menor velocidade do

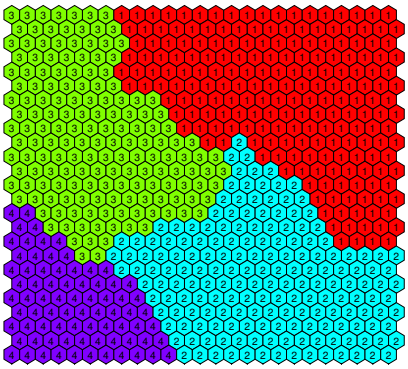

Figura 5. Divisão dos neurônios do SOM em 4 *cluters* determinados pela análise hierárquica.

Tabela III
MÉDIA DOS VALORES DAS VARIÁVEIS DE ACORDO COM OS CLUSTERS FORMADOS PELA REDE SOM.

Variáveis	Média das variáveis por cluster			
	1	2	3	4
SO_2 (ppb)	0,18	0,09	0,17	0,89
CO (ppb)	153,18	126,03	443,43	230,86
O_3 (ppb)	11,93	7,38	5,45	7,61
MP (μg/m3)	17,73	12,92	17,97	15,83
NO (ppb)	9,20	9,15	29,64	19,28
NO_2 (ppb)	6,10	6,81	12,29	9,10
Velocidade do vento (m/s)	4,15	1,83	2,26	2,20
Temperatura (°C)	26,44	24,10	24,80	23,97
Umidade (%)	64,30	76,97	73,15	74,38
STWD (°)	21,59	26,16	26,39	23,43
Amostras	5.240	4.469	3.799	2.027

vento média, temperatura mediana e maior umidade relativa. Sendo um grupo caracterizado por sua baixa concentração geral de poluentes com 29% dos dados.

As maiores concentrações de CO, MP, NO e NO_2 estão no *cluster* 3 da Tabela III, as exceções são o SO_2 e O_3 com baixos valores, tendo o O_3 a menor média total. A velocidade do vento, temperatura e umidade relativa possuem valores intermediários. Com 24% dos dados atribuídos, o *cluster* 3 é caracterizado por valores altos de concentração dos poluentes.

Por fim, como último *cluster* da Tabela III o *cluster* 4, principalmente caracterizado pela maior e considerável concentração do poluente SO_2 comparado com os demais. Os outros poluentes apresentam valores de concentração intermediários, assim como a velocidade do vento, temperatura e umidade relativa. Com 2.027 amostras atribuídas (13%), o *cluster* 4 possui a menor quantidade de amostras.

C. Correlação entre as Variáveis

Os planos de componentes permitem uma ideia inicial e preliminar sobre as variáveis, que pode ser evidenciada por uma análise de correlação [17]. Os vetores de pesos dos neurônios obtidos pelo treinamento da rede SOM, também conhecido como *codebook*, retratam a aproximação e o conhecimento obtido sobre as variáveis, e podem ser utilizados para sua representação. Utilizando o *codebook*, a Fig. 6 apresenta a similaridade entre as variáveis empregando o critério de *Ward* e o coeficiente de correlação de *Pearson*, r.

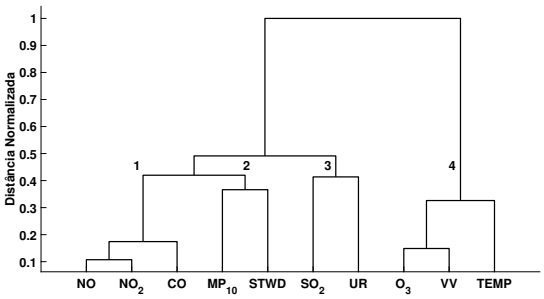

Figura 6. Correlação entre todas as variáveis utilizando o critério de *Ward* e distância baseada na similaridade de $1-r$, onde r é o coeficiente de *Pearson*.

Duas separações são bem visíveis na análise de correlação (Fig. 6). Porém, para uma análise mais adequada foram identificados neste experimento 4 grupos, sendo os três primeiros mais próximos entre si e que incluem todos os poluentes, com exceção do O_3, cuja origem é exclusivamente fotoquímica, e um último agrupamento contendo ozônio, velocidade do vento e temperatura. A Fig. 6 ilustra a correlação entre as variáveis e os 4 grupos identificados numerados de 1 a 4.

No grupo 1, os poluentes NO, NO_2 e CO possuem uma forte similaridade, o que pode indicar a possibilidade de compartilharem de uma mesma fonte de emissão, possivelmente veicular, devido a especificação de alocação da estação, assim como da região de monitoramento. Os mesmos juntamente relacionados com MP_{10}, que possivelmente também possui origem veicular e foi agrupado ao STWD, que é um indicativo da estabilidade atmosférica, quanto maior, menos estável a atmosfera e mais facilitada a dispersão e diluição dos poluentes. A correlação do MP_{10} com STWD no grupo 2 pode ser um indício de que o material particulado não está somente sendo produzido pela frota, mas como também dispersado pela movimentação existente na via de tráfego por conta do desgaste das vias e das pastilhas de freios dos veículos.

O SO_2 e a UR juntos no grupo 3 evidencia como a UR pode influenciar no processo de formação/decomposição de moléculas durante processos heterogêneos (fase líquida). O SO_2 em especial, pode reagir juntamente com a umidade do ar e outros oxidantes na atmosfera formando o ácido sulfúrico H_2SO_4, como também o sulfato de amônio.

O O_3 sendo um poluente secundário mostra ser bastante influenciado por parâmetros meteorológicos como a velocidade do vento [32], observado aqui no grupo 4 e nos planos de componentes. O que se pode avaliar é a possibilidade que talvez o O_3 não possua sua produção no local de monitoramento, e que na verdade seja transportado pela influência dos ventos juntamente com outros possíveis poluentes como os COVs. A temperatura também pode vir a ser outro possível responsável, já que maiores temperaturas aumentam a velocidade dos processos químicos, possibilitando a formação do ozônio.

IV. Conclusões

Nesse trabalho foi utilizada a técnica de aprendizagem de máquina mapas auto-organizáveis para análise de dados de poluentes atmosféricos. Foram utilizados dados fornecidos pelo Governo do Estado da Bahia das concentrações de poluentes e dados meteorológicos do período que compreendeu os anos de 2011 a 2016, da estação Itaigara localizada na cidade de Salvador - Bahia. A análise dos dados aplicando a rede SOM permitiu, a partir dos seus planos de componentes, determinar preliminarmente a identificação de padrões e correlações entre as variáveis estudadas. Além da utilização dos planos de componentes, foi possível realizar um balanço das características inerentes às amostras por meio da obtenção de *clusters* formados a partir do agrupamento dos neurônios da rede, contanto também com técnicas auxiliares. Fazendo uso de uma análise de correlação entre os vetores de pesos da saída da rede SOM, que representam as variáveis dos poluentes estudados, pôde-se evidenciar particularidades de como esses poluentes se relacionam e como influenciam em suas formações e possíveis fontes de emissão.

Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Pro-reitoria de Pesquisa, Pós-graduação e Inovação do IFBA (PRPGI-IFBA) pelo suporte e financiamento.

Referências

[1] P. J. Landrigan, R. Fuller, N. J. Acosta, O. Adeyi, R. Arnoldet, N. Basu, and et. al, "The lancet commission on pollution and health," *The Lancet*, vol. 391, no. 10119, p. 462–512, 2017.

[2] J. G. Zivin and M. Neidell, "Air pollution's hidden impacts," *Science*, vol. 359, no. 6371, pp. 39–40, 2018.

[3] M. C. Turner, Z. J. Andersen, Andrea, W. R. Diver, S. M. Gapstur, C. A. Pope III, D. Prada, j. Samet, G. Thurston, and A. Cohen, "Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 6, pp. 460–479, 2020.

[4] J. Zhang, L. Zhang, M. Du, W. Zhang, X. Huang, Y. Zhang, Y. Yang, J. Zhang, S. Deng, F. Shen, Y. Li, and H. Xiao, "Indentifying the major air pollutants base on factor and cluster analysis, a case study in 74 chinese cities," *Atmospheric Environment*, vol. 144, pp. 37–46, 2016.

[5] K. Zhang and S. Batterman, "Air pollution and health risks due to vehicle traffic," *Science of The Total Environment*, vol. 450-451, pp. 307–316, 2013.

[6] L. Bai, j. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4, pp. 307–316, 2018.

[7] D. Núñez-Alonso, L. V. Pérez-Arribas, S. Manzoor, and J. O. Cáceres, "Statistical tools for air pollution assessment: Multivariate and spatial analysis studies in the madrid region," *Journal of Analytical Methods in Chemistry*, vol. 2019, pp. 1–9, 2018.

[8] D. Tian, J. Fan, H. Jin, H. Mao, D. Geng, S. Hou, P. Zhang, and Y. Zhang, "Characteristic and spatiotemporal variation of air pollution in northern china based on correlation analysis and clustering analysis of five air pollutants," *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 8, pp. 1–12, 2020.

[9] P. Manimaran and A. C. Narayana, "Multifractal detrended cross-correlation analysis on air pollutants of university of hyderabad campus, india," *Physica A: Statistical Mechanics and its Applications*, vol. 502, pp. 228–235, 2018.

[10] Y. Bai, X. Jin, X. Wang, X. Wang, and J. Xu, "Dynamic correlation analysis method of air pollutants in spatio-temporal analysis," *International Journal of Environmental Research and Public Health*, vol. 17, no. 1, 2020.

[11] S. Zhao, Y. Yu, D. Yin, J. He, N. Liu, J. Qu, and J. Xiao, "Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from china national environmental monitoring center," *Environment International*, vol. 86, pp. 92–106, 2016.

[12] D. Yin, S. Zhao, and J. QU, "Spatial and seasonal variations of gaseous and particulate matter pollutants in 31 provincial capital cities, china," *Air Quality, Atmosphere & Health*, vol. 10, no. 3, p. 359–370, 2016.

[13] C. Li, Z. Wang, B. Li, Z. Peng, and Q. Fu, "Investigating the relationship between air pollution variation and urban form," *Building and Environment*, vol. 147, pp. 559–568, 2019.

[14] N. Periš, M. Buljac, M. Bralić, M. Buzuk, S. Brinić, and I. Plazibat, "Characterization of the air quality in split, croatia focusing upon fine and coarse particulate matter analysis," *Analytical Letters*, vol. 48, no. 3, pp. 553–565, 2015.

[15] C. Wang, L. Zhao, W. Sun, J. Xue, and Y. Xie, "Identifying redundant monitoring stations in an air quality monitoring network," *Atmospheric Environment*, vol. 190, pp. 256–268, 2018.

[16] R. Zhi-Yong and H. Bao-Gang, "Parameter identifiability in statistical machine learning: A review," *Neural Computation*, vol. 29, no. 5, pp. 1151–1203, 2017.

[17] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin: Springer-Verlag, 2001.

[18] U. Asan and S. Ercan, *An Introduction to Self-Organizing Maps*. Paris: Atlantis Press, 2012, pp. 295–315.

[19] J. L. Pearce, L. A. Waller, H. H. Chang, M. Klein, J. A. Mulholland, J. A. Sarnat, S. E.Sarnat, M. J. Strickland, and P. E. Tolbert, "Using self-organizing maps to develop ambient air quality classifications: a time series example," *Environmental Health*, vol. 11, no. 1, p. 56, 2014.

[20] B. Zhong, L. Wang, T. Liang, and B. Xing, "Pollution level and inhalation exposure of ambient aerosol fluoride as affected by polymetallic rare earth mining and smelting in baotou, north china," *Atmospheric Environment*, vol. 167, pp. 40–48, 2017.

[21] N. Jiang, Y. Scorgie, M. Hart, M. L. Riley, J. Crawford, P. J. Beggs, G. C. Edwards, L. Chang, D. Salter, and G. D. V. , "Visualising the relationships between synoptic circulation type and air quality in sydney, a subtropical coastal-basin environment," *International Journal of Climatology*, vol. 37, no. 3, pp. 1211–1228, 2017.

[22] V. Moosavi, G. Aschwanden, and E. Velasco, "Finding candidate locations for aerosol pollution monitoring at street level using a data-driven methodology," *Atmospheric Measurement Techniques*, vol. 8, no. 9, pp. 3563–3575, 2015.

[23] S. Kwon, W. Jeong, D. Park, k. Kim, and K. H. Cho, "A multivariate study for characterizing particulate matter (pm10, pm2.5, and pm1) in seoul metropolitan subway stations, korea," *Journal of Hazardous Materials*, vol. 297, pp. 295–303, 2015.

[24] F. Chang, L. Chang, C. Kang, Y. Wang, and A. Huang, "Explore spatio-temporal pm2.5 features in northern taiwan using machine learning techniques," *Science of The Total Environment*, vol. 736, p. 139656, 2020.

[25] D. Li and Y. Liao, "Pollution zone identification research during ozone pollution processes," *Environmental Monitoring and Assessment*, vol. 192, no. 9, p. 591, 2020.

[26] L. P. L. Fávero and P. P. Belfiore, *Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata*, 1st ed. Rio de Janeiro: Elsevier, 2017.

[27] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1358–1384, 1996.

[28] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.

[29] G. Pölzlbauer, "Survey and comparison of quality measures for self-organizing maps," in *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*. Elfa Academic Press, 2004, pp. 67–82.

[30] K. Kiviluoto, "Topology preservation in self-organizing maps," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 1, 1996, pp. 294–299 vol.1.

[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[32] D. Li and Y. Liao, "Pollution zone identification research during ozone pollution processes," *Environmental Monitoring and Assessment*, vol. 192, no. 591, 2020.