18º Congresso Interinstitucional de Iniciação Científica – CIIC 2024
27, 28 e 29 de agosto de 2024
ISSN: 2965-2812

IDENTIFICAÇÃO DE ZONAS DE RISCO DE ENTRADA DA PRAGA *XANTHOMONAS ORYZAE PV. ORYZAE* POR MEIO DE DIFERENTES TÉCNICAS DE CLUSTERIZAÇÃO DE DADOS ORDINAIS

Paulina Kayse de Andrade **Santos**[1]; Marcos Aurélio Santos da **Silva**[2]; Leonardo Nogueira **Matos**[3]; Gastão Florêncio **Miranda Júnior**[3]; Rafael **Mingoti**[4]; Cristiaini **Kano**[5]; Márcia Helena Galina **Dompieri**[6]

Nº 24506

RESUMO *– Esta pesquisa objetivou identificar zonas de risco de entrada da praga Xanthomonas oryzae* pv. *oryzae no Brasil, a partir de fatores de risco como: presença de portos e aeroportos; importação de frutos hospedeiros; vias e volume de transporte de carga e pessoas; proximidade a centros urbanos; proximidade com países onde a praga é presente e áreas de fronteira com dispersão ativa. Foram aplicados três métodos de clusterização de dados ordinais, baseados na análise de correlação múltipla combinada com o k-médias, na segmentação do mapa auto-organizável, e na segmentação da tabela de contingência multidirecional. Os resultados foram comparados com a superfície de risco elaborada a partir da média ponderada das classes ordinais, consideradas como valores intervalares. Dentre as zonas de risco, os métodos baseados no mapa auto-organizável e na análise de correlação múltipla obtiveram melhor desempenho. A partir da análise da assimetria para cada variável-agrupamento, foi possível identificar que as variáveis relacionadas ao transporte aéreo de pessoas e de cargas, à proximidade com regiões com registro da praga e ao ingresso por dispersão ativa estão mais associadas ao risco de entrada. Por sua vez, as variáveis relacionadas aos municípios com importação de hospedeiros de países em que a praga está presente e à proximidade às áreas urbanas são as que menos estão associadas ao risco de entrada, o que pode indicar ajustes sobre os pesos de cada variável no cálculo da média ponderada, presente no método de geração de superfície de risco.*

Palavras-chaves: análise de correlação múltipla, mapa auto-organizável, redes neurais artificiais, tabela de contingência multidirecional.

1 Autor, Bolsista CNPq (PIBIC): Graduação em Ciência da Computação, Universidade Federal de Sergipe, Aracaju-SE; paulina.santos@colaborador@embrapa.br.
2 Colaborador: Pesquisador da Embrapa Tabuleiros Costeiros, Aracaju-SE.
3 Colaboradores: Professores da Universidade Federal de Sergipe, Aracaju-SE.
4 Colaborador: Analista da Embrapa Territorial, Campinas-SP.
5 Colaboradora: Pesquisadora da Embrapa Territorial, Campinas-SP.
6 Orientadora: Pesquisadora da Embrapa Territorial, Campinas-SP; marcia.dompieri@embrapa.br.

ISSN: 2965-2812

IDENTIFICATION OF ENTRY RISK ZONES FOR THE XANTHOMONAS ORYZAE PV. ORYZAE PEST USING DIFFERENT ORDINAL DATA CLUSTERING TECHNIQUES

ABSTRACT – *This research aimed to identify areas at risk of entry of the Xanthomonas oryzae* pv. *oryzae pest in Brazil, based on risk factors such as: the presence of ports and airports; import of host fruits; routes and volume of transportation of cargo and people; proximity to urban centers; proximity to countries where the pest is present and bordering areas with active dispersal. We applied three ordinal data clustering methods based on Multiple Correlation Analyses combined with k-means, Self-Organizing Map segmentation, and multidirectional contingency table segmentation. We compared the results with the risk surface prepared from the ordinal classes' weighted average, which are considered as interval values for these thematic maps. Among the risk zones, the methods based on the Self-Organizing Map and the Multiple Correlation Analysis obtained better results. From the analysis of asymmetry for each grouping variable, we determined that the variables related to air transport of people and cargo, proximity to regions with a record of the plague, and entry through dispersion are the ones most associated with entry risk. The variables related to municipalities importing hosts from countries where the pest is present and proximity to urban areas are least related with entry risk, which may indicate adjustments to the weights of each variable when calculating the weighted average present in the risk surface generation method.*

Keywords: Multiple Correlation Analysis, Self-Organizing Map, Artificial Neural Networks, Multidirectional Contingency Table.

1. INTRODUÇÃO

Pragas quarentenárias podem gerar grande impacto econômico para a agricultura e devem ser monitoradas de forma a evitar sua entrada no Brasil (pragas ausentes) ou a diminuir seus efeitos nos sistemas produtivos caso estejam presentes no território. O Ministério da Agricultura, Pecuária e Abastecimento (Mapa), em conjunto com outras instituições federais, estaduais e municipais, desenvolve estratégias para o monitoramento das principais pragas ausentes e presentes. Uma dessas estratégias envolve a geração de mapas de possíveis rotas de entrada e estabelecimento de determinada praga no país. Essas rotas são elaboradas a partir de oficinas e grupos focais que estudam os fatores de entrada, de dispersão e de estabelecimento, para que medidas fitossanitárias sejam aplicadas no país.

Dentre as pragas monitoradas, destaca-se a praga quarentenária *Xanthomonas oryzae* pv. *oryzae* (Xanthomonadales: Xanthomonadaceae). Conhecida popularmente como crestamento ou murcha bacteriana, é responsável por uma das mais importantes e devastadoras doenças do arroz, seu principal hospedeiro, mas pode atingir também gramíneas e espécies de arroz silvestre.

ISSN: 2965-2812

As possíveis rotas de risco para a praga *Xanthomonas oryzae* pv. *oryzae* foram elaboradas a partir do mapa de superfície de risco de entrada da praga, obtido a partir da média ponderada das classes ordinais de risco para cada uma das variáveis que compõem o fator risco de entrada (Silva et al., 2023a).

O objetivo desta pesquisa foi gerar zonas de risco de entrada a partir do mesmo conjunto de dados usados por Silva et al. (2023b), usando três diferentes técnicas de clusterização de dados ordinais, identificando vantagens, desvantagens e complementaridades entre elas e em relação ao método usado para construção da superfície de risco. Foram avaliados: o método de clusterização das projeções a partir da análise de correlação múltipla (duas dimensões com as maiores inércias) usando o método k-médias (Ranalli; Rocci, 2015); o método de clusterização a partir da segmentação do mapa auto-organizável (Silva et al., 2023b); e o método baseado na segmentação da tabela de contingência multidirecional (Giordan; Diana, 2011).

2. MATERIAL E MÉTODOS

2.1 Superfície de risco de entrada da *Xanthomonas oryzae* pv. *oryzae* no Brasil

A superfície de risco de entrada da *Xanthomonas oryzae* pv. *oryzae* no Brasil (Figura 1) foi elaborada no contexto de determinação das principais rotas de entrada e estabelecimento da praga a partir da média ponderada de sete variáveis definidas através de oficinas de trabalho e grupos focais (Silva et al., 2023a). Cada variável corresponde a um mapa temático cujas classes assumem os valores inteiros entre zero e dez. A classe “zero” denota que naquela região não há risco de entrada da praga, a classe “dez” representa o risco máximo de entrada, e valores intermediários significam risco crescente entre zero e dez. As variáveis selecionadas foram: portos e aeroportos com importação de hospedeiros de países em que a praga está presente (PORTIMP, peso 7,5); municípios com importação de hospedeiros de países em que a praga está presente (MUNIMP, peso 7,5); transporte aéreo de pessoas e de carga (TRANSAC, peso 15); trânsito de pessoas e cargas (TRANSTP, peso 15); proximidade a áreas urbanas (PROXURB, peso 15); ingresso por dispersão ativa pela fronteira (INGRDISP, peso 10); proximidade a regiões com registro da praga (PROXPRG, peso 20).

2.2 Clusterizando dados categóricos ordinais

Para as próximas seções, para n observações há J variáveis ordinais categóricas, $J \geq 1$, representadas por $Y = Y_1, \ldots, Y_j$, e N_j denota o número de categorias ou modalidades para $Y_j, j = 1, \ldots, J$.

ISSN: 2965-2812

2.2.1 Análise de agrupamentos de dados ordinais com base na transformação do dado

A estratégia mais simples para a clusterização de dados ordinais é a sua transformação para valores intervalares, para que seja possível a aplicação das técnicas tradicionais como as que usam distância euclidiana como k-médias, hierárquico aglomerativo/divisivo, DBSCAN, etc. Uma estratégia possível é a aplicação do método de análise de correspondência múltipla sobre a tabela Burt e a posterior aplicação do algoritmo k-médias sobre as projeções dos indivíduos considerando as duas dimensões com maiores inércias (Ranalli; Rocci, 2015). Essa técnica tem como principais vantagens a simplicidade e o baixo custo computacional, que permitem sua aplicação a dados massivos. No entanto, as relações de proximidade no espaço gerado pelas duas dimensões podem não corresponder à real estrutura do dado.

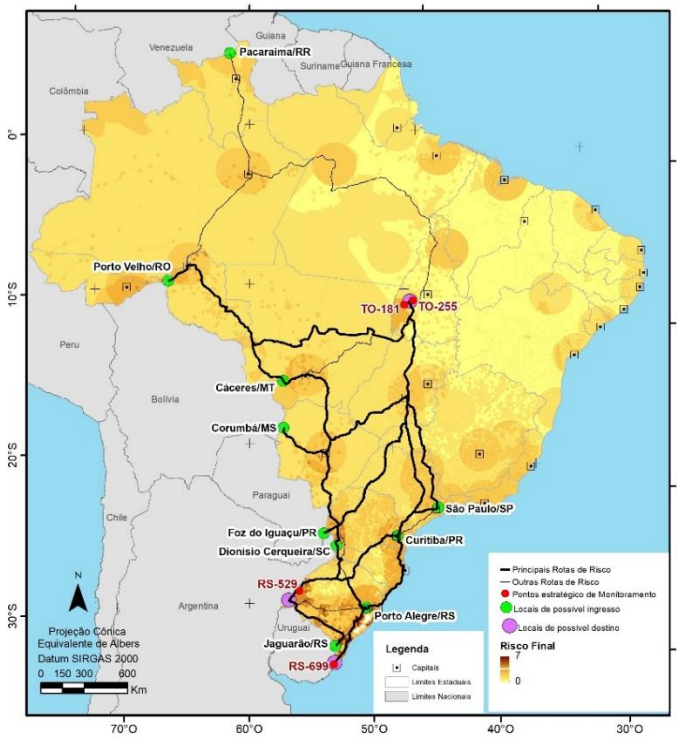


Figura 1. Mapa de superfície do risco de entrada no Brasil da praga quarentenária *Xanthomonas oryzae* pv. *oryzae* elaborado a partir da média ponderada dos planos de informação do risco de entrada.
Fonte: Silva et al. (2023a).

2.2.2 Agrupamento com base em limiarização da tabela de contingência multidirecional

O algoritmo proposto em Giordan e Diana (2011) foi projetado explicitamente para dados categóricos ordinais e usa a tabela de contingência multidirecional do próprio conjunto de dados como ponto de partida. Cada célula dessa tabela representa observações com as mesmas características, portanto elas devem fazer parte do mesmo *cluster*. Além disso, como estamos lidando com dados ordinais, as células vizinhas podem implicar alguma proximidade entre as observações associadas a cada uma delas. Assim, os autores usam essa ideia de proximidade entre as células para propor um algoritmo de agrupamento que, primeiro, considera a densidade da célula

ISSN: 2965-2812

medida como uma frequência ou proporção de observações e, em seguida, considera a vizinhança da célula para mesclar grupos ou associar células não rotuladas em um agrupamento.

De acordo com os autores, a posição de uma célula C em uma tabela de contingência multidirecional é dada por $(C_1,\ldots,C_j)$, de modo que $C_j = 1,\ldots,N_j$, então a vizinhança de C é o conjunto de células cujas coordenadas são $(I_1,\ldots,I_j) \in Int(1)\times\ldots\times Int(j)\backslash(C_1,\ldots,C_j)$, em que (Equação 1):

$$Int(j) = \begin{cases} \{1,2\}, & se\ C_j = 1 \\ \{N_j - 1, N_j\}, & se\ C_j = N_j \\ \{C_j - 1, C_j, C_j + 1\}, & caso\ contrário \end{cases} \quad (1)$$

A partir disso, podemos descrever o método proposto por Giordan e Diana (2011, p. 1318) a partir do Algoritmo 1. Esse método tem baixo custo computacional, não necessita transformar o dado ordinal em intervalar, e tem um único parâmetro a ser definido empiricamente, a constante lambda.

Algoritmo 1 Agrupamento da tabela de contingência multidirecional

Requer: P — Tabela de contingência normalizada multidirecional contendo as proporções p para cada célula.
Requer: $\lambda \in [0,1]$ — Limite para determinar os clusters iniciais

1. Localizar as células com alta proporção p, $p \geq \lambda$
2. Atribuir células vizinhas com alta proporção p ao mesmo cluster
3. **enquanto** não há células com $0 < p < \lambda$ na vizinhança das rotuladas **faça**
4. **se** a célula não está rotulada e $0 < p < \lambda$ e tem apenas um vizinho rotulado **então**
5. essa célula assume o rótulo da vizinha
6. **senão**
7. essa célula assume um rótulo de ruído
8. **fim-se**
9. **fim-enquanto**
10. As células não rotuladas serão rotuladas como ruído

2.3 Método baseado na segmentação do mapa auto-organizável (MAO)

O método proposto por Silva et al. (2023b) trata os dados ordinais como dados numéricos intervalares. O dado categórico ordinal é transformado de modo que a distância entre duas classes subsequentes seja a mesma para todas as variáveis, dividindo esses valores pelo valor mais elevado de todas as variáveis (no nosso caso, o maior valor possível é a modalidade 10). Assim, nosso vetor de atributos Y' terá sete componentes que variam seus valores entre 0/10 e 10/10, ou seja, no intervalo [0,1].

O MAO é uma rede neural artificial (RNA) com aprendizado de máquina não supervisionado, no qual os neurônios artificiais são representados por vetores de peso, w, com a mesma dimensão dos dados de entrada. Eles são organizados em uma grade bidimensional, $N \times M$, com uma estrutura hexagonal que define a vizinhança entre os neurônios. O mecanismo de aprendizado de máquina estocástico é iterativo e visa a atualização dos pesos w conforme a Equação 2 (Kohonen, 2001).

ISSN: 2965-2812

$$w(w+1) = w(t) + \alpha(t)h(t)(Y_i' - w(t)) \qquad (2)$$

t representa a iteração, $w(t)$ é o vetor de peso do neurônio na iteração t, $\alpha(t)$ é um pequeno valor que representa a taxa de aprendizado, $h(t)$ é uma função de vizinhança e Y_i' um vetor de dados de entrada tomado aleatoriamente. Ao fim das iterações, cada dado de entrada é associado a um único neurônio, que pode representar mais de um vetor de entrada. Os pesos do MAO preservam a topologia dos dados, o que significa que os neurônios vizinhos podem representar vetores de entrada próximos. Esse recurso do MAO permite executar algoritmos de agrupamento (por exemplo, k-*means* ou k-médias) sobre os pesos da rede neural como uma forma indireta de particionar os dados de entrada (Silva et al., 2022).

Entretanto, é possível segmentar o MAO sem o auxílio de algoritmos de agrupamento tradicionais e usando informações internas da rede neural, como distância e vizinhança entre os pesos do MAO, nível de ativação dos neurônios (número de vetores de entrada associados a ele) e densidade de dados entre os neurônios (Silva et al., 2023b).

Silva et al. (2023b) propuseram um algoritmo de segmentação baseado na interpretação do MAO como um grafo não direcionado, que usa todas as informações disponíveis após o processo de aprendizado de máquina, sendo necessário apenas definir o número desejado k de clusters (Algoritmo 2). A qualidade da partição e escolha definitiva do valor k pode ser auxiliada pelo uso do índice de validação de agrupamentos CDbw estabelecido por Halkidi e Varzigianis (2002), usando os pesos da RNA MAO treinada como vetores de referência do algoritmo conforme proposto por Silva et al. (2011).

Algoritmo 2: Método proposto por Silva, M.A.S. et al. (2023) utilizando MAO

Requer: G = (V,E) - Grafo do MAO treinado
Requer: H - Dados do nível de atividade dos neurônios
Requer: D - Matriz de distância entre os pesos
Requer: k - O número de clusters desejados

1. T <- árvore geradora mínima (MST) de G usando D como os pesos das arestas
2. **parar** cada aresta $(u,v)\ \epsilon\ T$ **faça**
3. $custo(u,v) \leftarrow DBI(u,v)$
4. **fim-para**
5. Podar as $k-1$ arestas em T com os menores custos
6. Atribuir um rótulo de cluster a cada conjunto de nós conectados em T.

A Tabela 1 mostra os resultados do agrupamento de alguns dados artificiais rotulados de referência (duplo espiral, gaussianas, duas correntes e íris) usando o k-médias, o DBSCAN, hierárquico aglomerativo, MAO + k-médias, MAO + hierárquico aglomerativo e o método proposto por Silva et al. (2023b). Observamos que o método proposto tem bom desempenho em todos os quatro conjuntos de dados. Avaliamos diferentes hiperparâmetros (número de neurônios e malha da

ISSN: 2965-2812

grade) para a rede neural artificial MAO, e usamos a medida de acurácia (ACC) para escolher a configuração final.

Tabela 1. Acurácia da clusterização dos dados usados para comparação dos diferentes métodos em associação ao método húngaro de associação entre rótulos dos grupos e os rótulos dos dados.

	K-médias	DBSCAN	Hierárquico aglomerativo	MAO + hierárquico aglomerativo	MAO + k-médias	Segmentação da MAO
Duplo espiral	0,64	0,99	0,67	0,59	0,61	1,00
Gaussianas	0,96	0,62	0,94	0,93	0,97	0,96
Duas correntes	0,51	1,00	0,5	0,71	0,65	1,00
Iris	0,83	0,68	0,83	0,88	0,83	0,87

2.3 Avaliação da qualidade dos agrupamentos e identificação dos grupos com maior risco

Como aplicamos diferentes estratégias de clusterização sobre dados não rotulados, decidimos avaliar a qualidade da clusterização observando, para cada variável, se o método dividiu os dados em grupos diferenciáveis com moda única e frequências decaindo em torno da moda conforme proposto por Biernacki e Jacques (2016). Para identificar os grupos com maior risco de entrada da praga, procedemos à análise da curva de densidade de distribuição de probabilidade por variável e por grupo a partir do coeficiente de Fisher para a assimetria. Como as categorias são ordinais e variam entre zero e dez, quanto menor e negativo for o valor da assimetria, maior será o risco naquele grupo para aquela variável. As análises foram feitas no ambiente python 3.11 usando as bibliotecas scikit-learn versão 0.23 e MiniSom versão 2.3.2.

3. RESULTADOS E DISCUSSÃO

Para cada um dos três métodos de clusterização de dados categóricos ordinais, geramos os histogramas para cada variável-grupo e um mapa com os grupos que contêm pelo menos uma variável com assimetria negativa, destacando os grupos com assimetria negativa mais proeminente (maior risco de entrada da praga) (Tabela 2).

3.1 Zoneamento do risco de entrada a partir da tabela Burt (ACM + k-médias)

Para este método, procedemos com a análise de correspondência múltipla (ACM) a partir da tabela Burt, e posteriormente aplicamos o algoritmo k-médias sobre as projeções das observações nas duas dimensões com maior inércia, avaliando diferentes valores de k, tendo definido k = 8 em função de o aumento do número de *clusters* não melhorar a diferenciação dos mesmos. A Figura 2a mostra coerência com o mapa de superfície da Figura 1, com concentração das áreas de maior risco na região litorânea entre o Sudeste e o Sul do Brasil, assim como na região da tríplice fronteira. A partir da Tabela 2, o risco de entrada da praga do grupo 2 está associado às variáveis “portos e aeroportos

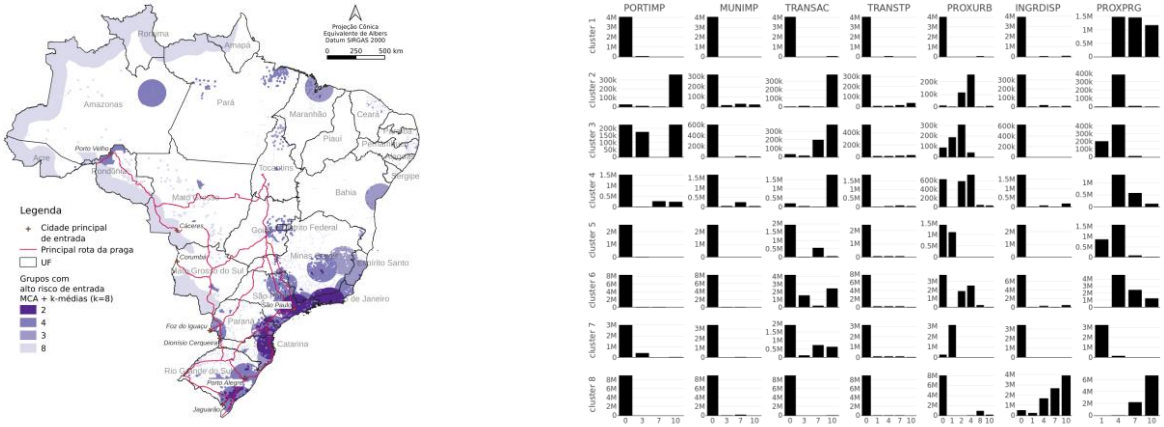
ISSN: 2965-2812

com importação de hospedeiros” e “transporte aéreo de pessoas e de carga”, esta última também responsável pelo risco nos grupos 3 e 4; o grupo 8 apresenta o menor risco entre os grupos em destaque, determinado pelas variáveis “proximidade a regiões com registro da praga” e “ingresso por dispersão ativa pela fronteira” (Figura 2b).

Tabela 2. Valores da assimetria da curva de densidade de probabilidade e percentual de observações (perc) para cada variável e grupo *c*, considerando os três métodos de clusterização de dados ordinais.

Método	c	perc (%)	PORTIMP	MUNIMP	TRANSAC	TRANSTP	PROXURB	INGRDISP	PROXPRG
MCA + k-médias (2 dimensões com maiores inércias)	1	48,0172	10,0	1396,0	9,6	14,5	8,0	5,7	0,1
	2	2,3151	-3,0	2,1	-5,3	2,0	1,1	4,2	3,6
	3	2,4452	0,4	5,3	-1,9	2,8	0,5	300,6	-0,5
	4	5,9755	1,3	2,0	-2,4	3,5	1,0	2,5	1,2
	5	5,9570	27,5	491,7	1,3	6,0	0,3	469,3	-0,1
	6	16,3845	20,0	7,8	0,6	5,2	0,9	3,3	0,8
	7	5,7109	3,1	96,1	0,5	4,2	-3,0	459,5	4,4
	8	13,1945	731,8	11,0	102,5	23,2	2,5	-0,8	-1,2
Segmentação da tabela de contingência (limiar = 0,00022)	1	0,1072	17,0	3,6	-3,5	-0,3	1,5	62,9	0,2
	2	1,2396	5,9	4,0	-4,1	4,2	1,4	214,1	0,5
	3	0,0240	10,8	27,0	10,7	5,0	11,2	-8,7	24,6
	4	16,3963	3,3	14,2	0,8	7,6	0,9	8,3	-0,3
	5	0,0741	54,8	0,2	47,4	5,9	50,1	-0,7	-8,0
	6	2,1971	2,8	6,6	0,8	0,3	1,0	3,5	0,4
	7	79,0320	4,9	8,2	2,0	10,9	3,0	1,9	-0,2
	8	0,3550	1,8	0,8	0,7	3,0	0,8	1,9	0,6
	9	0,0918	61,0	52,8	61,0	58,2	55,7	0,8	-60,8
	10	0,4829	121,2	140,0	6,0	13,8	3,9	-4,8	-6,1
Segmentação do mapa auto-organizável baseada em grafos	1	1,4027	1,8	9,5	2,7	-0,3	0,8	6,4	0,1
	2	0,2601	-6,4	2,2	-11,7	-0,6	0,1	9,1	-0,9
	3	43,0606	5,3	9,3	0,8	8,4	1,7	6,2	1,5
	4	4,0880	-1,9	2,8	-14,2	7,5	1,1	3,5	0,0
	5	5,0150	55,2	451,2	451,2	6,1	-0,5	430,6	449,5
	6	0,9416	195,5	10,2	-195,5	5,4	1,2	-6,7	-1,1
	7	15,7015	7,7	4,7	6,5	9,9	1,8	-0,9	-1,2
	8	16,6909	823,1	32,5	823,1	18,8	9,1	10,3	-649,5
	9	12,8396	96,1	14,2	721,9	13,1	70,7	9,6	-719,2

ISSN: 2965-2812

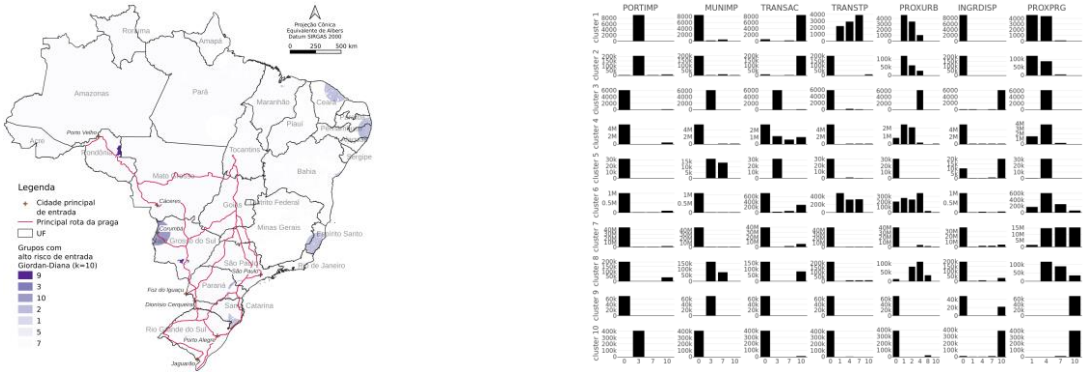


(a) (b)

Figura 2. Grupos de risco de entrada (a) e histogramas (b) gerados pelo algoritmo MCA + k-médias (k=8) para as variáveis (PORTIMP, MUNIMP, TRANSAC, TRANSTP, PROXURB, INGRDISP e PROXPRG).

3.2 Zoneamento do risco de entrada a partir do algoritmo Giordan-Diana

Avaliamos diferentes valores para o hiperparâmetro lambda ($2{,}2.10^{-3}$, $2{,}3.10^{-3}$ e $2{,}4.10^{-3}$) do algoritmo proposto por Giordan e Diana (2011), que geraram diferentes agrupamentos de risco de entrada da praga no Brasil, variando o número *k* de grupos para cada um desses valores, 10, 9 e 7, respectivamente. Para a análise, decidimos pela solução com o maior número de agrupamentos diferenciáveis. A Figura 3a mostra que o resultado obtido não se aproximou do mapa de superfície da Figura 1. No entanto, de acordo com a Tabela 2, as zonas de risco identificadas estão associadas às mesmas variáveis destacadas pelo método ACM + k-médias, exceto “portos e aeroportos com importação de hospedeiros” (Figura 3b).



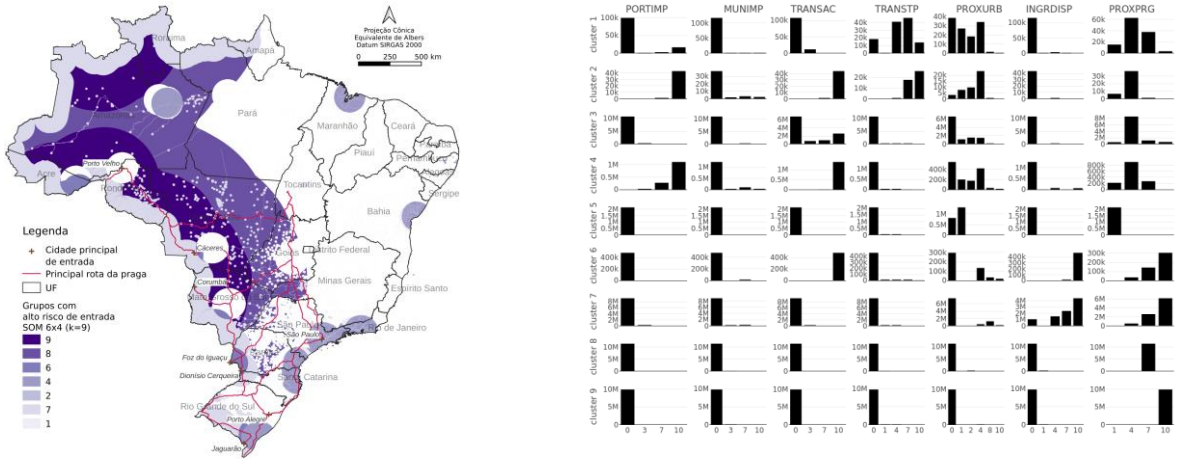
(a) (b)

Figura 3. Mapa dos grupos de risco de entrada (a) e histograma (b) gerados pelo algoritmo baseado em modelos (BOS) para as variáveis (PORTIMP, MUNIMP, TRANSAC, TRANSTP, PROXURB, INGRDISP e PROXPRG).

ISSN: 2965-2812

3.4 Zoneamento do risco de entrada a partir de RNA MAO

Foram avaliados diferentes tamanhos (NxM) de RNA MAO mantendo-se sempre a topologia planar uni e bidimensional, grade hexagonal, função de vizinhança gaussiana e aprendizado sequencial com iniciação randômica dos pesos da rede neural MAO. As diferentes configurações da RNA MAO foram avaliadas de acordo com erro de quantização e, após a segmentação pelo algoritmo proposto por Silva et al. (2023b), foi aplicado o índice CDbw (Halkidi; Varzigianis, 2002; Silva et al., 2011) usando os pesos da RNA MAO treinada como vetores de representatividade, para definir qual a melhor RNA para o conjunto de dados. Após esse processo, foi identificada a RNA MAO hexagonal 6x4 particionada em nove grupos (Figura 4a). A Figura 4a mostra que o método encontrou sete grupos de risco, compatível com o mapa de risco de entrada definido a partir da média ponderada da Figura 1. De acordo com a Tabela 2, as zonas de risco identificadas estão associadas às mesmas variáveis destacadas pelo método baseado na combinação do uso de ACM e k-médias, e, neste caso, os grupos com maiores assimetrias negativas (9 e 8) estão associados à variável “proximidade a regiões com registro da praga” (Figura 4b).



(a) (b)

Figura 4. Mapa dos grupos de risco de entrada (a) e histogramas (b) gerados pelo algoritmo proposto (RNA MAO 6x4 hexagonal) e histograma para as variáveis (PORTIMP, MUNIMP, TRANSAC, TRANSTP, PROXURB, INGRDISP e PROXPRG).

De acordo com a Tabela 2, as variáveis “municípios com importação de hospedeiros de países em que a praga está presente” (MUNIMP) e “proximidade a áreas urbanas” (PROXURB) não apresentaram nenhuma distribuição assimétrica negativa para nenhum método. De forma inversa, as demais variáveis apresentaram forte assimetria negativa para pelo menos um método. Os valores de assimetria indicam que há contribuições distintas de cada variável na composição do risco de entrada da praga e que os pesos atribuídos a cada uma para o cálculo da média ponderada proposto por Silva et al. (2023b) pode ser revisto.

O método baseado na segmentação da tabela de contingência multidirecional conseguiu identificar mais grupos, mas bem desbalanceados, e os grupos de risco corresponderam a uma área

bastante restrita. No entanto, mesmo não identificando zonas de risco perfeitamente compatíveis com a superfície de risco obtida por Silva et al. (2023b), esses métodos mostraram-se importantes para identificar as variáveis que mais contribuem para as áreas de risco, como complementares à análise, conforme mostraram Silva et al. (2023b).

4. CONCLUSÃO

Os resultados mostram que, a partir da análise dos histogramas, os métodos baseados na transformação dos dados (RNA MAO e MCA + k-médias) geraram grupos com moda e decaimento em torno da moda melhor definidos. Todos os métodos geraram zonas de risco compatíveis com a superfície de risco de entrada definida por Silva et al. (2023b), no entanto os métodos baseados na RNA MAO e no MCA + k-médias geraram grupos com melhor identificação das áreas de risco.

A análise dos valores das assimetrias permitiu identificar as variáveis que mais contribuíram para os riscos em cada grupo, e todos os métodos foram coerentes nessa identificação. Assim, podemos destacar as variáveis “transporte aéreo de pessoas e de carga” (TRANSAC), “proximidade a regiões com registro da praga” (PROXPG) e “ingresso por dispersão ativa pela fronteira” (INGRDISP) como as que mais estão relacionadas ao risco de entrada, e as variáveis “municípios com importação de hospedeiros de países em que a praga está presente” (MUNIMP) e “proximidade a áreas urbanas” (PROXURB) como as que menos estão associadas ao risco de entrada. Esse resultado pode indicar ajustes nos pesos de cada variável usada no cálculo da média ponderada na geração de superfície de risco.

5. AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio ao trabalho por meio do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI).

6. REFERÊNCIAS

BIERNACKI, C.; JACQUES, J. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. **Statistical Computing**, v. 26, p. 929–943, 2016. DOI: 10.1007/s11222-015-9585-2.

GIORDAN, M.; DIANA, G. A clustering method for categorical ordinal data. **Communications in Statistics—Theory and Methods**, v. 40, n. 7, p. 1315–1334, 2011. DOI: 10.1080/03610920903581010.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multirepresentatives. In: SETN CONFERENCE, 2., 2002, Thessaloniki. **Proceedings…** Greece, 2002.

KOHONEN, T. **Self-Organizing Maps**. Berlin: Springer, 2001.

RANALLI, M.; ROCCI, R. Clustering Methods for Ordinal Data: A Comparison between Standard and New Approaches. In: MORLINI, I.; MINERVA, T.; VICHI, M. (ed.). **Advances in Statistical Models for Data**

Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham. 2015. DOI: https://doi.org/10.1007/978-3-319-17377-1_23.

SILVA, B. H. S. A.; et al. **Mapeamento territorial estratégico de potenciais rotas de risco de dispersão da praga *Xanthomonas oryzae* pv. *oryzae***. Campinas, SP: Embrapa Territorial, 2023a.

SILVA, M. A. S.; MATOS, L. N.; SANTOS, F. E. de O.; DOMPIERI, M. H. G.; MOURA, F. R. de. Tracking the connection between Brazilian agricultural diversity and native vegetation change by a machine learning approach. **IEEE Latin America Transactions**, v. 20, n. 11, p. 2371–2380, 2022. DOI: 10.1109/TLA.2022.9904762.

SILVA, M. A. S.; BARRETO, P. V. de A.; MATOS, L. N.; MIRANDA JUNIOR, G. F.; DOMPIERI, M. H. G.; MOURA, F. R. de; RESENDE, F. K. S.; NOVAIS, P.; OLIVEIRA, P. A Self-Organizing Map Clustering Approach to Support Territorial Zoning. In: VASCONCELOS, V.; DOMINGUES, I.; PAREDES, S. (ed.). **Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications**. Heidelberg: Springer, 2023b. p. 289-303. DOI: https://doi.org/10.1007/978-3-031-49018-7_2.

SILVA, M. A. S.; SIQUEIRA, E. R. de; TEIXEIRA, O. A.; MANOS, M. G. L.; MONTEIRO, A. M. V. Using Self-Organizing Maps for Rural Territorial Typology. In: PRADO, H. A. do; LUIZ, A. J. B.; CHAIB FILHO, H. (org.). **Computational methods for agricultural research**: advances and applications. Hershey: Information Science Reference, 2011. v. 1, p. 107-126. DOI: 10.4018/978-1-61692-871-1.ch007.