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

Marcos A. S. da Siva[1], *Marcia G. H. Dompieri*[2], *Paulina K. A. Santos*[3], *Leonardo N. Matos*[3], *Gastão F. Miranda Júnior*[3], *Rafael Mingoti*[2] *e Cristiaini Kano*[2]

[1]Embrapa Tabuleiros Costeiros, Av. Gov. Paulo B de Menezes, 3250, Aracaju, SE, marcos.santos-silva@embrapa.br, [2]Embrapa Territorial, Campinas, SP {marcia.dompieri,rafael.mingoti,cristiaini.kano}@embrapa.br e [3]Universidade Federal de Sergipe, São Cristóvão, kaspaulina@hotmail.com, leonardo@dcomp.ufs.br, gastao@mat.ufs.br

RESUMO

Esta pesquisa teve como objetivo a avaliação de diferentes técnicas de identificação, via clusterização, de zonas de risco de entrada da praga *Xanthomonas Oryzae pv. oryzae* no Brasil, a partir de mapas temáticos com classes ordinais relacionados a fatores de risco. Foram aplicados métodos baseados na Análise de Correlação Múltipla/k-médias, na segmentação do Mapa Auto Organizável, e na limiarização da tabela de contingência. Os resultados foram comparados com a superfície de risco elaborada a partir da média ponderada das classes. Para avaliar a qualidade dos agrupamentos foram avaliados se para variável e grupo o histograma tinha uma única moda proeminente, com as frequências das demais modalidades decaindo em torno da moda e se os grupos eram diferenciáveis pelas modas. A análise da assimetria dos histogramas permitiu identificar grupos com maior risco. A rede neural apresentou o melhor resultado e as assimetrias dos histogramas permitiu identificar as variáveis associadas ao risco.

Palavras-chave – Análise de Correlação Múltipla, Mapa Auto Organizável, Redes Neurais Artificiais, Tabela de contingência.

ABSTRACT

This study aimed to evaluate different techniques for clustering areas at risk of entry of the pest Xanthomonas Oryzae pv. oryzae in Brazil, based on thematic maps with ordinal classes related to risk factors. We applied Methods based on Multiple Correlation Analysis/k-means, Self-Organizing Map segmentation, and contingency table thresholding. We compared the results with the risk surface developed from the weighted average of the classes. To assess the quality of the clusters, we evaluated whether the histogram had a single prominent mode for the variable and group, with the frequencies of the other modalities decreasing around the mode, and whether the groups were distinguishable by the modes. The analysis of the asymmetry of the histograms allowed us to identify groups with higher risk. The neural network presented the best result, and the asymmetries of the histograms allowed us to identify the variables associated with risk.

Key words – *Artificial Neural Networks, Contingency Table, Multiple Correlation Analysis, Self-Organizing Map.*

1. INTRODUÇÃO

Pragas quarentenárias podem gerar grande impacto econômico negativo para a agricultura e devem ser monitoradas de forma a se evitar sua entrada no Brasil, pragas ausentes, ou diminuir seus efeitos nos sistemas produtivos caso estejam presentes em nosso território. O Ministério da Agricultura, Pecuária e Abastecimento (MAPA), em conjunto com outras instituições federais, estaduais e municipais, desenvolve estratégias para o controle de ingresso e contingenciamento das principais pragas ausentes e presentes. Uma dessas envolve a geração de mapas de possíveis rotas de entrada e estabelecimento de determinada praga no país [1]. Essas rotas são elaboradas a partir de oficinas e grupos focais que estudam os fatores de entrada, de dispersão, de estabelecimento e impactos, para que medidas fitossanitárias sejam estabelecidas no país.

Dentre as pragas monitoradas destaca-se a praga quarentenária *Xanthomonas Oryzae pv. oryzae* (Xanthomonadales: Xanthomonadaceae), conhecida popularmente como crestamento ou murcha bacteriana, que 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. 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 compõe o fator risco de entrada [2].

O objetivo desta pesquisa foi avaliar três métodos de geração de zonas de risco de entrada a partir do mesmo conjunto de dados ordinais usados por [2], identificando aquele com melhor desempenho considerando a qualidade dos grupos. Foram avaliados o método de clusterização das projeções a partir do Análise de Correlação Múltipla (duas dimensões com as maiores inércias) usando o método k-médias [3], o método de clusterização a partir da segmentação do Mapa Auto-Organizável [4], e o método baseado na tabela de contingência multidirecional [5].

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 (Fig. 1) foi elaborada no contexto de determinação das principais rotas de entrada e

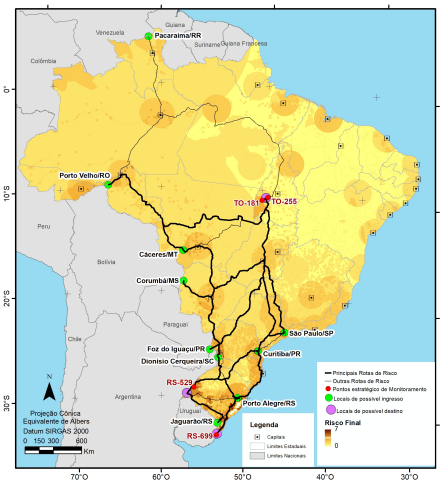

Figura 1: Mapa de superfície do risco de entrada no Brasil da praga quarentenária *Xanthomonas Oryzae pv. oryzae* elaborado a partir da média ponderada dos planos de informação do risco de entrada. Fonte: [2]

estabelecimento da praga a partir da média ponderada de sete variáveis definidas através de oficinas de trabalho e grupos focais [2]. Cada variável corresponde a um mapa temático cujas classes assumem os valores inteiros entre zero e dez. Sendo que 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) com peso 7,5; municípios com importação de hospedeiros de países em que a praga está presente (MUNIMP) com peso 7,5; transporte aéreo de pessoas e de carga (TRANSAC) com peso 15; trânsito de pessoas e cargas (TRANSTP) com peso 15; proximidade a áreas urbanas (PROXURB) com peso 15; ingresso por dispersão ativa pela fronteira (INGRDISP) com peso 10; proximidade a regiões com registro da praga (PROXPRG) com peso 20.

2.2. Clusterizando dados categóricos ordinais

2.2.1. Análise de agrupamentos de dados ordinais baseado na sua projeção linear

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 a 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 posteriormente aplicação do algoritmo k-médias sobre as projeções dos indivíduos considerando as duas dimensões com maiores inércias [3]. Essa técnica possui como principal vantagem a simplicidade e o baixo custo computacional, que permite sua aplicação a dados massivos. No entanto, as relações de proximidade no espaço gerado pelas duas dimensões podem não corresponder a real estrutura do dado. O conjunto pode ter uma estrutura não-linear que pode não ser capturada pelo MCA e pelo k-médias.

2.2.2. Agrupamento baseado na segmentação da tabela de contingência

O algoritmo proposto em [5] foi projetado explicitamente para dados categóricos ordinais e usa a tabela de contingência múltipla do próprio conjunto de dados como ponto de partida (Algoritmo 1). Cada célula desta 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 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 com base num limiar definido por meio de tentativa-e-erro.

Algorithm 1 Segmentação da tabela de contingência

Input: P ▷ Tabela de contingência normalizada múltipla contendo as proporções p para cada célula.
Input: $\lambda \in [0, 1]$ ▷ Limite para determinar os clusters iniciais.
Localizar as células com alta proporção p, $p \geq \lambda$
Atribuir células vizinhas com alta proporção p ao mesmo cluster
while não há células com $0 < p < \lambda$ na vizinhança das células rotuladas **do**
  if a célula não esta rotulada e $0 < p <$ e tem apenas um vizinho rotulado **then**
    essa célula assume o rótulo da vizinha
  else if N is odd **then**
    essa célula assume um rótulo de ruído
  end if
end while
As células não rotuladas serão rotuladas como ruído

2.2.3. Método baseado na segmentação do Mapa Auto Organizável

O método proposto por [4] é baseado na rede neural MAO e foi concebida originalmente para dados intervalares. Logo, foi necessário transformar os dados ordinais em 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 Mapa Auto-Organizável é uma rede neural artificial com aprendizado não supervisionado, onde 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, NM, 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 de forma a termos um mapeamento topológico entre o conjunto de dados de entrada e os pesos da rede neural [6]. Isto significa que os neurônios vizinhos podem representar vetores de entrada próximos, e que dados próximos estarão associados a neurônios vizinhos. Esse recurso permite aplicar algoritmos de agrupamento (e.g., k-means) sobre os pesos da rede neural como uma forma indireta de particionar os dados de entrada [7]. Entretanto, é possível segmentar a rede MAO sem o auxílio de algoritmos de agrupamento tradicionais 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. A partir deste princípio, Silva et al. [4] propuseram o Algoritmo 2.

Algorithm 2 Segmentação da rede neural Mapa Auto-Organizavel

Input: $G = (V, E)$ ▷ Mapa auto-organizável após o processo de Aprendizado de Máquina.
Input: D ▷ Matriz de distância entre os pesos da rede neural.
Input: k ▷ O número de clusters desejados
$T \leftarrow$ árvore geradora mínima (AGM) de G usando D como os pesos das arestas
for cada aresta $(u, v) \in T$ **do**
$custo(u, v) \leftarrow DBI(u, v)$
end for
Podar as $k - 1$ arestas em T com os menores custos
Atribuir um rótulo de cluster a cada conjunto de nós conectados em T

2.2.4. Comparação entre os algoritmos e determinação das zonas de risco

Para comparação entre os três métodos de clusterização definimos três critérios de qualidade de acordo com [8] e baseados nos histogramas por variável e grupo gerados: número de histogramas onde há uma moda proeminente, número de histogramas onde as demais classes em torno da moda apresentam frequências que decaem em torno dela, e se os agrupamento são diferenciáveis a partir das modas.

Para avaliar quais grupos apresentam áreas de maior risco avaliamos a assimetria dos histogramas. Quanto mais acentuada e negativa é a assimetria maior o risco de entrada da região representada por aquele grupo.

3. RESULTADOS E DISCUSSÃO

A Tabela 1 resume as informações dos histogramas por variável e grupo obtidos a partir da segmentação dos dados ordinais pelas três técnicas avaliadas, MCA + k-médias (Fig. 2), limiarização da tabela de contingência (Fig. 3) e segmentação da rede neural MAO (Fig. 4). Observa-se que o método baseado na rede neural obteve os melhores resultados de forma que todas as variáveis por grupo apresentam moda proeminente, apenas oito histogramas não apresentam as demais classes decaindo em torno da moda e todos os grupos são diferenciáveis.

Tabela 1: Resultados dos critérios de avaliação da qualidade da clusterização de dados ordinais com base na existência de moda proeminente com demais classes com frequência decaindo no seu entorno e diferenciabilidade dos grupos pelas modas.

Método	A	B	C
MCA + k-médias	1	18	Sim
Limiarização da tabela de contigência	2	11	Sim
Segmentação do MAO	0	8	Sim

A: Nº de histogramas sem moda proeminente.
B: Nº de histogramas cujas demais classes no entorno da moda não possuem frequências que decaem em torno dela.
C: Grupos diferenciáveis pela moda.

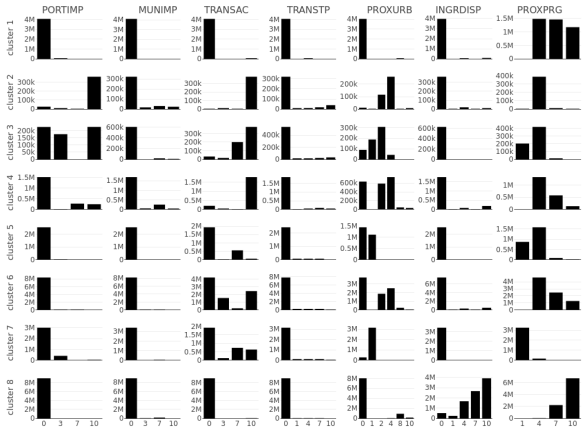

Figura 2: Histogramas gerados pelo algoritmo MCA + k-médias

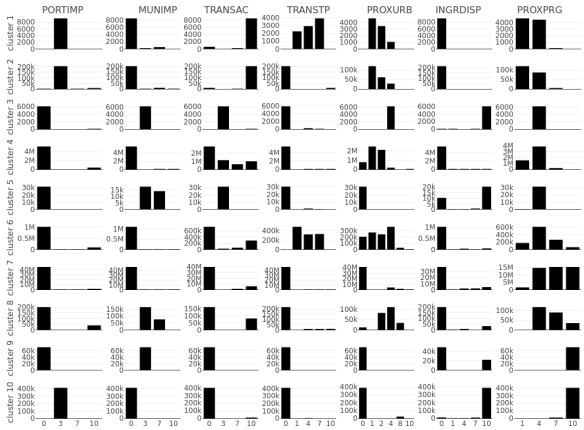

Figura 3: Histogramas gerados pelo algoritmo de segmentação da tabela de contingência

A Tabela 2 mostra os valores das assimetrias para todos os histogramas da Fig. 4. A partir desses valores foi possível determinar quais os grupos (9, 8, 6, 4, 2, 7 e 1, em ordem decrescente das assimetrias) apresentam valores que representam áreas com alto risco de entrada da praga no Brasil.

A análise dos valores das assimetrias permitiu identificar as variáveis que mais contribuíram para os riscos em cada grupo, sendo que todos os métodos foram coerentes nessa

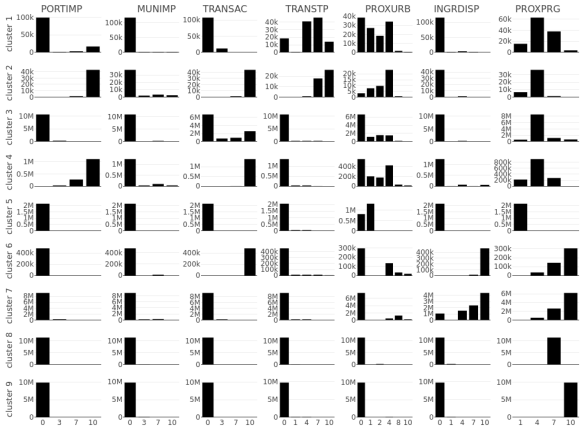

Figura 4: Histogramas gerados pela segmentação do Mapa Auto-Organizável (RNA 2D 6x4 hexagonal)

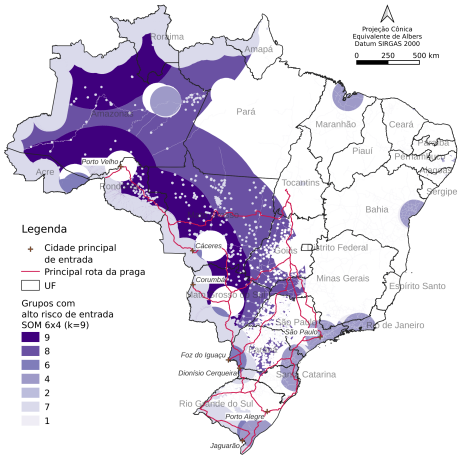

Figura 5: Mapa dos grupos de risco de entrada gerados pelo algoritmo proposto (RNA MAO 6x4 hexagonal)

identificação. Assim, a partir da partição dos dados usado a rede neural podemos destacar as variáveis transporte aéreo de pessoas e de carga (TRANSAC), proximidade a regiões com registro da praga (PROXPG), 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 no cálculo da média ponderada na geração de superfície de risco.

Tabela 2: Assimetrias para cada variável e grupo c gerado pelo método de clusterização baseado na rede neural MAO e segmentação baseada em grafos

c	V1	V2	V3	V4	V5	V6	V7
1	1,8	9,5	2,7	-0,3	0,8	6,4	0,1
2	**-6,4**	2,2	**-11,7**	**-0,6**	0,1	9,1	-0,9
3	5,3	9,3	0,8	8,4	1,7	6,2	1,5
4	**-1,9**	2,8	**-14,2**	7,5	1,1	3,5	0,0
5	55,2	451,2	451,2	6,1	-0,5	430,6	449,5
6	195,5	10,2	**-195,5**	5,4	1,2	**-6,7**	**-1,1**
7	7,7	4,7	6,5	9,9	1,8	**0,9**	**-1,2**
8	823,1	32,5	823,1	18,8	9,1	10,3	**-649,5**
9	96,1	14,2	721,9	13,1	70,7	9,6	**-719,2**

V1 (PORTIMP). V2 (MUNIMP). V3 (TRANSAC). V4 (TRANSTP). V5 (PROXURB). V6 (INGRDIP). V7 (PROXPRG).

4. CONCLUSÕES

A análise realizada demonstrou de maneira contundente a eficácia da segmentação por meio da rede neural na identificação de padrões de risco relacionados à entrada de pragas. O método se destacou por proporcionar uma representação eficaz da maioria das variáveis com moda proeminente e assegurar a diferenciabilidade entre os grupos. Essa capacidade de segmentação é fundamental, pois possibilita uma compreensão mais refinada dos dados, permitindo que decisões informadas sejam tomadas. Os resultados evidenciaram variáveis críticas que estão fortemente associadas ao risco de entrada, como o TRANSAC, a PROXPG e o INGRDISP. Essas variáveis devem ser priorizadas em qualquer estratégia de monitoramento e controle, pois representam os maiores pontos de vulnerabilidade. Em contraste, variáveis como MUNIMP e a PROXURB mostraram-se menos relevantes, sugerindo que esforços focados nessas áreas podem ser menos eficazes. O mapa de grupos de risco gerado pelo algoritmo proporciona a visualização das áreas de maior vulnerabilidade, o que é crucial para a formulação de estratégias proativas de mitigação. Com base nesses insights, recomenda-se que políticas e ações de controle sejam desenvolvidas considerando as variáveis identificadas como de maior impacto, permitindo uma alocação eficiente de recursos e uma resposta mais ágil às ameaças de entrada de pragas.

5. REFERÊNCIAS

[1] R Mingoti et al. Territorial zoning of brazilian areas favorable to anastrepha curvicauda (diptera: Tephritidae) in papaya cultivation. *J. of Agric. Sci. Res.*, 2:2–10, 2022.

[2] MHG Dompieri et al. Mapeamento territorial estratégico de potenciais rotas de risco de dispersão da praga *Xanthomonas oryzae pv. oryzae*. Technical report, Embrapa Territorial, 2023.

[3] M. Ranalli and R. Rocci. Clustering methods for ordinal data: A comparison between standard and new approaches. In I. Morlini, T. Minerva, and M. Vichi, editors, *Adv. in Stat. Models for Data Anal.*, volume 1, Cham, 2015. Springer.

[4] M. A. S. da Silva et al. A self-organizing map clustering approach to support territorial zoning. In V. Vasconcelos, I. Domingues, and S. Paredes, editors, *CIARP 2023. Lecture Notes in Computer Science.*, volume 1, pages 272–286, Cham, 2023. Springer.

[5] M. Giordan and G. Diana. A clustering method for categorical ordinal data. *Commun. Stat.- Theory Methods*, 40(7):1315–1334, 2011.

[6] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, 2001.

[7] MAS da Silva et al. Tracking the connection between Brazilian agricultural diversity and native vegetation change by a machine learning approach. *IEEE Lat. Am. T.*, 20(11):2371–2380, 2022.

[8] C. Biernacki and J. Jacques. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Stat. Comput.*, 26:929–943, 2016.