Revista Brasileira de Computação Aplicada, Julho, 2024

DOI: 10.5335/rbca.v16i2.9999
Vol. 16, Nº 2, pp. 31–44
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Simulação da qualidade de cafés utilizando critérios da ABIC na discriminação de regiões com diferentes estruturas de dependência espacial

Simulation of coffee quality using ABIC criteria in discriminating regions with different spatial dependency structures

Laryssa Ribeiro Calcagnoto [1], and Marcelo Ângelo Cirillo [1]

[1]Universidade Federal de Lavras

*laryssa.calcagnoto1@estudante.ufla.br; macufla@ufla.br

Recebido: 12/12/2023. **Revisado:** 07/07/2024. **Aceito:** 31/07/2024.

Resumo

O artigo aborda a aplicação da análise discriminante em dados de proporção com dependência espacial, com foco na discriminação de cafés especiais em microrregiões simuladas. Utilizando simulações Monte Carlo, o estudo explora cenários com diferentes níveis de dependência espacial e funções de ligação, comparando os efeitos de modelos de semivariograma (esférico, exponencial e gaussiano) sobre a acurácia e precisão das análises. Os resultados indicam que a análise discriminante quadrática tende a ser mais eficaz para amostras menores (n = 25), enquanto a dependência espacial forte favorece o modelo exponencial, melhorando a especificidade e reduzindo a taxa de falso positivo com o aumento de n. Ao empregar a função de ligação complemento log-log, padrões semelhantes são observados, com o modelo gaussiano se destacando em amostras maiores. O estudo destaca a importância de escolher o modelo apropriado com base em fatores como dependência espacial, tamanho da amostra e função de ligação para garantir resultados precisos. O equilíbrio entre especificidade e sensibilidade é crucial na decisão final, evidenciando a necessidade de uma abordagem criteriosa para a análise de dados de proporção com dependência espacial.

Palavras-Chave: Análise Discriminante; Café; Dependência Espacial; Simulação Monte Carlo.

Abstract

The paper addresses the application of discriminant analysis in proportion data with spatial dependence, focusing on the discrimination of specialty coffees in simulated microregions. Using Monte Carlo simulations, the study explores scenarios with varying levels of spatial dependence and link functions, comparing the effects of semivariogram models (spherical, exponential, and Gaussian) on the accuracy and precision of the analyses. The results indicate that quadratic discriminant analysis tends to be more effective for smaller samples (n = 25), while strong spatial dependence favors the exponential model, improving specificity and reducing the false positive rate as n increases. When employing the complementary log-log link function, similar patterns are observed, with the Gaussian model performing best in larger samples. The study highlights the importance of selecting the appropriate model based on factors such as spatial dependence, sample size, and link function to ensure accurate results. The balance between specificity and sensitivity is crucial in the final decision, emphasizing the need for a careful approach to the analysis of proportion data with spatial dependence.

Keywords: Coffee; Discriminant Analysis; Monte Carlo simulation; Spatial Dependence.

1 Introdução

Em função da complexidade em analisar dados sensoriais, o uso de metodologia e modelos estatísticos tem sido amplamente empregado. Na análise da discriminação da qualidade do café, por exemplo, de Oliveira et al. (2023) utilizaram o algoritmo *Expectation-Maximization* (E.M) para avaliar a performance do nível sensorial. Consideraram dois grupos de provadores, os quais discriminaram cafés especiais produzidos na região da Serra da Mantiqueira, com diferentes altitudes.

No tocante ao uso de modelos, Manoel et al. (2024) utilizaram técnicas de simulação para validar o modelo hierárquico adaptativo na viabilidade de discriminar a qualidade de cafés especiais armazenados em câmaras de resfriamento. Concluíram que o modelo é adequado para as considerações experimentais adotadas.

Com o objetivo de classificar ou diferenciar observações provenientes de duas ou mais populações, as técnicas de análise discriminante são extensivamente empregadas e continuamente objeto de pesquisa e avaliação para aprimoramento. Estas técnicas demonstram uma notável capacidade de adaptação a diversas situações, incorporando recursos computacionais para otimizar sua eficácia e precisão.

A técnica estatística de análise discriminante, conforme discutido por Johnson and Wichern (2007) e Hair Jr. et al. (2019), é fundamentada na presença de amostras em uma escala contínua, onde as variáveis medidas são tratadas como contínuas. Esta abordagem tem como objetivo a classificação ou discriminação entre distintos grupos, fundamentando-se nas características multivariadas dessas variáveis.

Neste contexto, Rausch and Kelley (2009) conduziram um estudo com o objetivo de comparar a acurácia da classificação proveniente da análise discriminante com outros procedimentos, como a regressão logística, considerando a não normalidade dos dados. Concluiu que a análise discriminante produziu melhores taxas de precisão, especialmente quando a distribuição dos dados apresenta um elevado grau de simetria.

Em se tratando do aprimoramento desta técnica, Bouveyron et al. (2007) formalizaram a adaptação da análise discriminante para altas dimensões, reduzindo a dimensão em cada classe de forma independente, com a vantagem de reduzir o número de parâmetros a serem estimados.

Reportando o uso da técnica de análise discriminante aprimorada com a incorporação de machine learning, mais especificamente o método de boosting (Faria et al. (2016); Menezes et al. (2016)), em aplicações que envolvem a qualidade de cafés na análise sensorial, Oliveira et al. (2019) propuseram uma análise de um experimento sensorial relacionado a testes de aceitação com cafés especiais. Considerou dois grupos de provadores previamente definidos, diferenciados pela presença e ausência de treinamento.

Com essas especificações, em comparação com os métodos convencionais, os autores observaram que as taxas de acerto foram superiores, resultando na redução das taxas de falso negativo. Posteriormente, de Oliveira et al. (2023), para a mesma aplicação, consideraram multiclasses definidas pelos grupos de provadores treinados com idade entre 19 e 50 anos e provadores não treinados com idades acima de 50 anos, concluiu que a taxa de erro global foi reduzida em comparação com os resultados obtidos pela análise discriminante linear e quadrática.

O intuito deste artigo é conduzir análises de simulação, explorando a distribuição espacial por meio de diversos modelos de semivariogramas para discriminar a qualidade dos cafés. Estas avaliações serão pautadas pelos critérios estipulados pela Associação Brasileira da Indústria de Café (ABIC).

2 Contextualização do problema como objeto de estudo via simulação Monte Carlo

O objeto de estudo proposto neste artigo aborda a conjectura relacionada à produção de cafés em duas microrregiões distintas. Em cada ponto de coleta, é empregado um sistema de referência que inclui latitudes e longitudes, representando diversas localidades e/ou lavouras.

Para cada uma dessas localidades, uma seleção de amostras é realizada, submetendo-as a uma avaliação sensorial. As notas finais, atribuídas em uma escala contínua de 0 a 1, são indicativas da qualidade intrínseca do café.

A avaliação de qualidade, nesse contexto, segue rigorosamente os critérios de classificação estipulados pela Associação Brasileira da Indústria de Café (ABIC), conforme representado na Fig. 1.

Esses padrões garantem que os cafés submetidos à avaliação atendam aos requisitos necessários para serem reconhecidos como aceitáveis e adequados ao consumo.

A adesão a tais normas não apenas assegura a qualidade do produto final, mas também promove a transparência e a confiabilidade ao longo de toda a cadeia de produção e comercialização de café, beneficiando consumidores, produtores e a indústria como um todo.

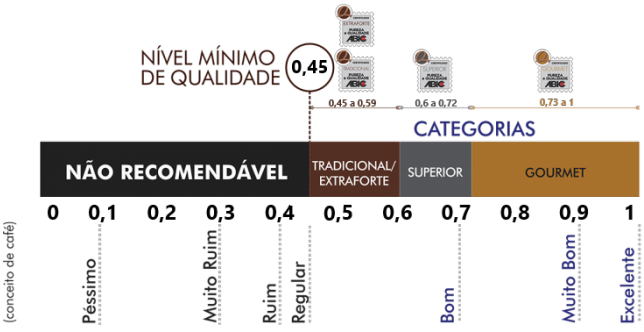


Figura 1: Nível da Qualidade do Café em Uma Escala de 0 a 1
Fonte: Adaptado de ABIC (2020)

Com base no exposto, a Tabela 1 apresenta as variáveis a serem incorporadas no estudo de simulação. A tabela serve como um guia fundamental, identificando e descrevendo claramente os fatores considerados, estabelecendo as bases para a modelagem e execução da simulação.

Tabela 1: Descrição das variáveis a serem incorporadas no procedimento de simulação

Variáveis	Natureza	Representação
X, Y	Regionalizados	X: Latitude; Y: Longitude
m_1, m_2	Binárias	Identificação das microrregiões m_1 e m_2.
N	Contínua na escala de proporção $0-1$	Nota final obtida na avaliação sensorial de cada amostra.
Q	Binárias	Qualidade mínima exigida pela ABIC, sendo $Q=1$ se $N \geq 0,45$ e $Q=0$ se $N<0,45$.
K	Contagem	Quantidade de amostras avaliadas em cada localidade.

Fonte: Autores (2024).

3 Metodologia

3.1 Simulação e Implementação dos modelos

Na execução do procedimento de simulação, em função do problema apresentado (Seção 2), foram gerados inicialmente dois grids regulares de dimensões $p \times p$, com $p=5$ e $p=10$, no espaço bidimensional representado por $v=(x,y)$. A geometria foi definida por células quadráticas.

Com a obtenção das coordenadas mencionadas em $v=(x,y)$, o processo de simulação das notas finais, nas escalas de $0-1$, foi realizado, considerando os modelos generalizados especificados para média (μ) a dados de proporção. Neste contexto, assume-se o preditor $\eta = X+Y$ com as funções de ligação logit Eq. (1) e complemento log-log Eq. (2).

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \eta \tag{1}$$

$$g(\mu) = \log(-\log(1-\mu)) = \eta \tag{2}$$

Dessa forma, as médias preditas, para cada função de ligação considerando a estrutura de dependência espacial foram geradas pelos modelos.

$$\mu = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$\mu = \exp\{-\exp(1-\eta)\}$$

Assim, as notas finais (N) foram geradas considerando $N \sim \text{Binomial}(n, \mu)$, onde $n=K$ representa o número de ocorrências, sendo geradas com a probabilidade μ obtida pelos modelos Logístico e complemento log-log.

Nessas condições, nos grids gerados, foram integrados diferentes modelos de semivariogramas em cenários de fraca e forte dependência espacial, como resumido pelos índices IDE definidos em Eq. (3).

$$IDE = \frac{\tau^2}{\tau^2+\sigma^2} \tag{3}$$

τ^2 é interpretado como a variação espacial não explicada, sendo justificado por escalas menores do que a distância entre os pontos amostrados, enquanto σ^2 é a variação espacial explicada pela autocorrelação espacial.

Com o valor calculado do IDE, é possível classificar o grau de dependência espacial em uma tabela, seguindo as categorias apropriadas.

Tabela 2: Classificação do Índice de Dependência Espacial (IDE)

IDE	Classificação
Acima de 75%	Forte
Entre 25% e 75%	Médio
Abaixo de 25%	Fraco

Fonte: Adaptado de Cambardella et al. (1994).

Logo, os parâmetros utilizados nos semivariogramas e associados aos grids encontram-se descritos na Tabela 3.

Tabela 3: Descrição dos parâmetros dos modelos de semivariograma, associados aos grids, considerando as respostas das notas geradas pelos modelos logit e complemento log-log (cloglog)

Cenário	Função de Ligação	Modelo Semivariograma	Parâmetros
IDE Forte - 0,80			
1	Logit	Esférico	$\tau^2=1,5$; $\sigma^2=0,37$
2		Gaussiano	$\tau^2=1,5$; $\sigma^2=0,37$
3		Exponencial	$\tau^2=1,5$; $\sigma^2=0,37$
4	clog-log	Esférico	$\tau^2=1,5$; $\sigma^2=0,37$
5		Gaussiano	$\tau^2=1,5$; $\sigma^2=0,37$
6		Exponencial	$\tau^2=1,5$; $\sigma^2=0,37$
IDE Fraco - 0,23			
7	Logit	Esférico	$\tau^2=1,5$; $\sigma^2=5$
8		Gaussiano	$\tau^2=1,5$; $\sigma^2=5$
9		Exponencial	$\tau^2=1,5$; $\sigma^2=5$
10	clog-log	Esférico	$\tau^2=1,5$; $\sigma^2=5$
11		Gaussiano	$\tau^2=1,5$; $\sigma^2=5$
12		Exponencial	$\tau^2=1,5$; $\sigma^2=5$

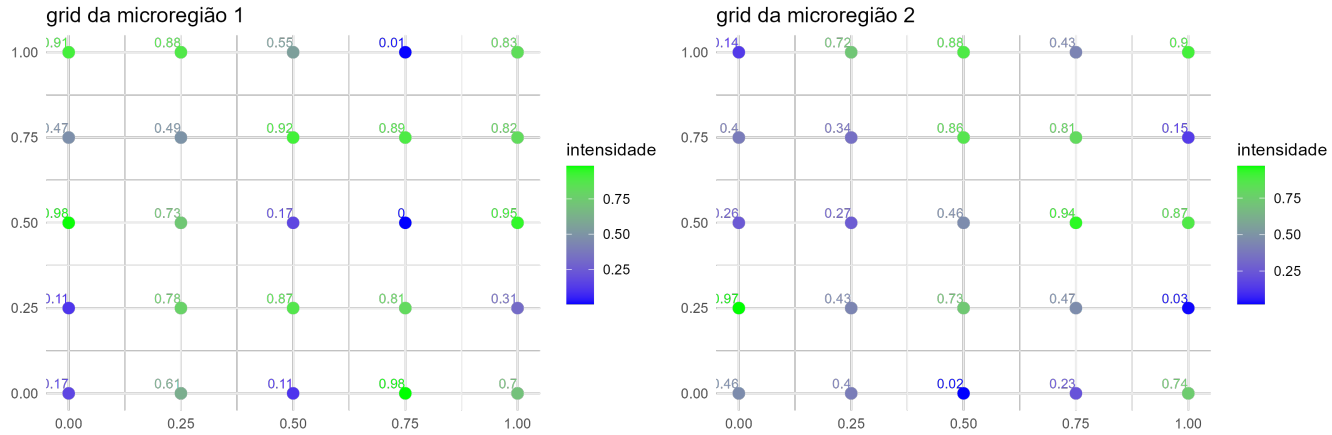
Fonte: Autores (2023).

Dessa forma, é apresentado um layout dos grids gerados para situações de fraca e forte dependência espacial, conforme indicado nas Tabelas 4 e 5.

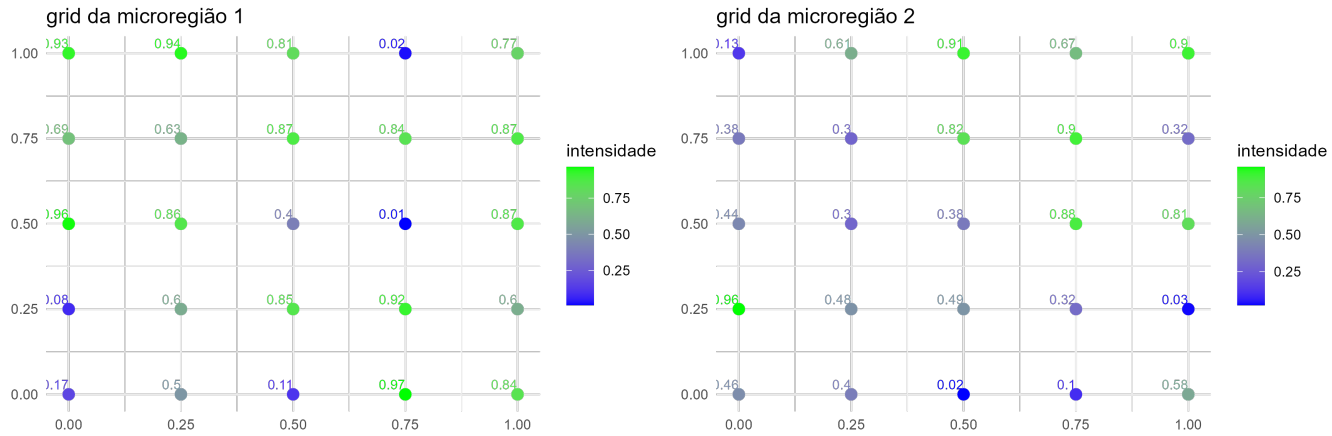
Analogamente, foram gerados grids de dimensão 10×10 para as funções de ligação, juntamente com grids 5×5 para a função complemento log-log. Os pontos estão dispostos na escala cromática variando de azul a verde. Os tons mais intensos de verde são associados a notas mais altas, enquanto os próximas do azul são notas mais baixas.

Tabela 4: Layout da distribuição de pontos amostrais considerando n = 25 com IDE = 0,23 quando utilizado função de ligação logit.

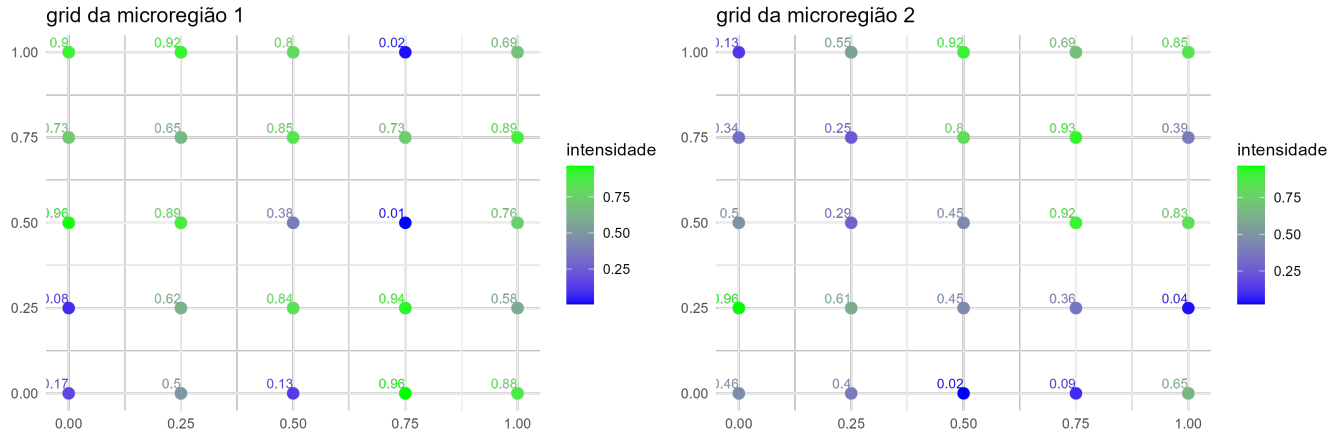
Esférico



Exponencial



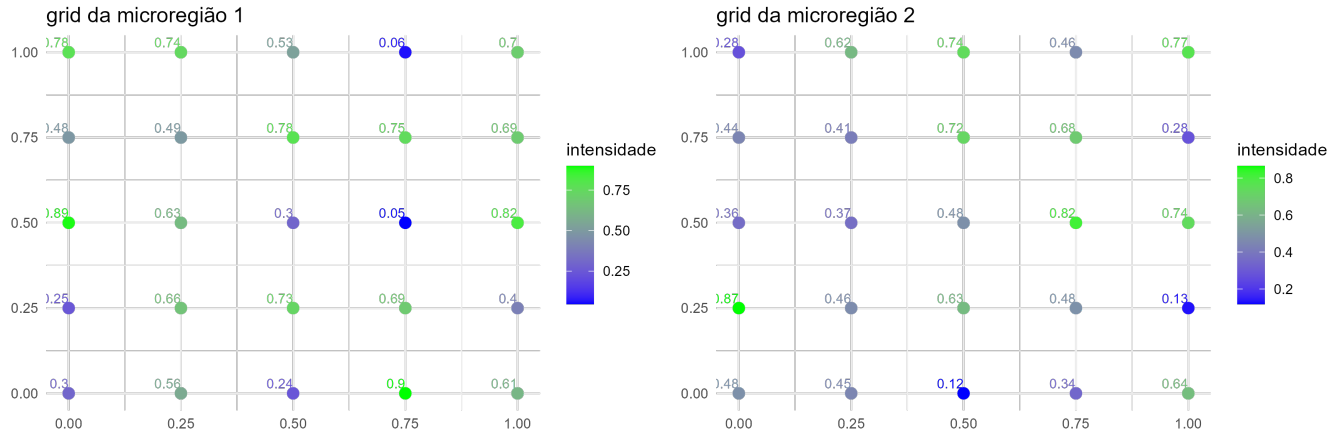
Gaussiano



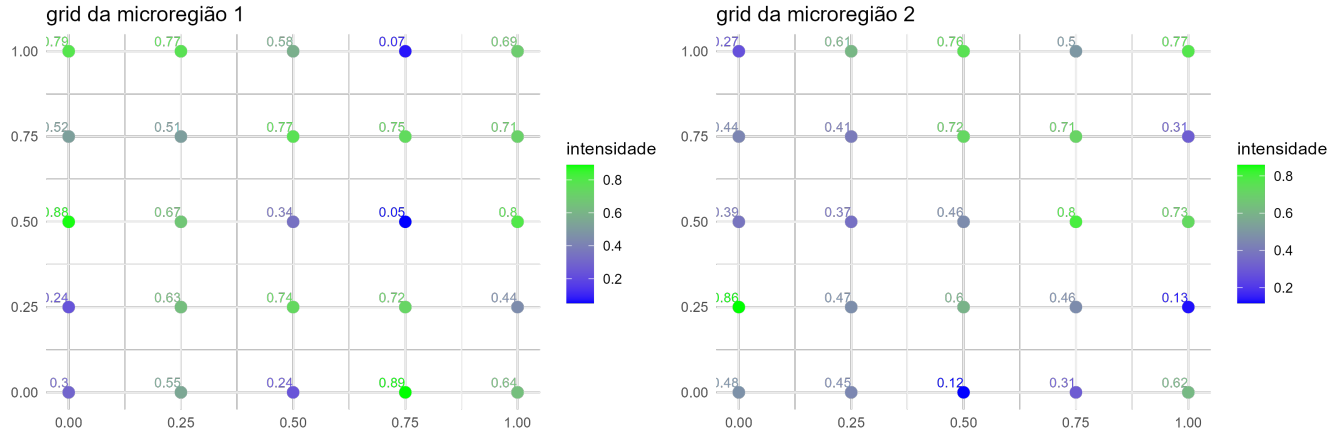
Fonte: Os autores (2023).

Tabela 5: Layout da distribuição de pontos amostrais considerando n = 25 pontos experimentais com IDE = 0,80 quando utilizado função de ligação logit.

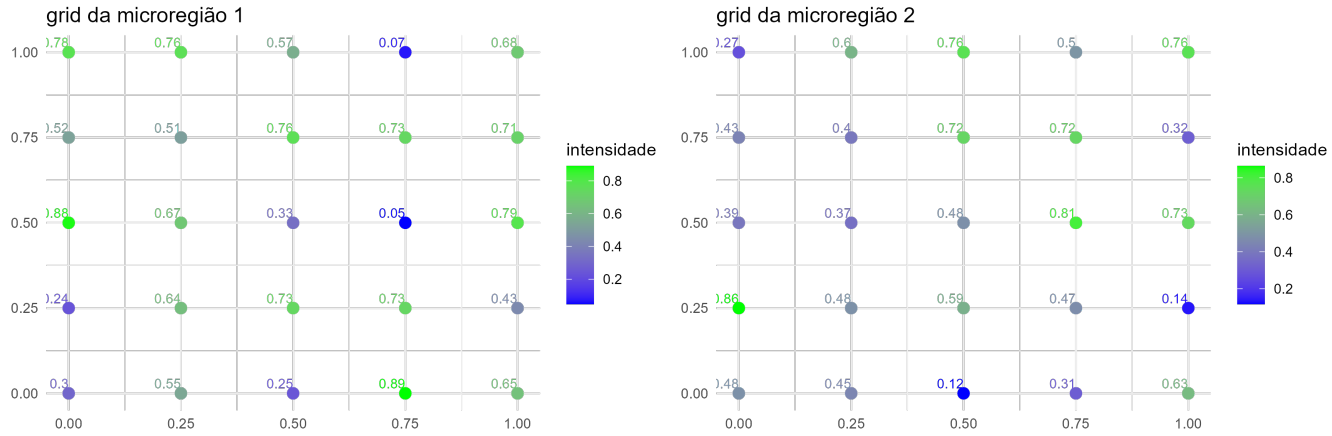
Esférico



Exponencial



Gaussiano



Fonte: Os autores (2023).

3.2 Viabilidade da aplicação de análise discriminante e medidas de desempenho em relação a classificação das microrregiões simuladas

Para cada cenário (conforme apresentado na Tabela 3), em conformidade com a suposição de homogeneidade entre as matrizes de covariância necessárias na análise discriminante linear de Fisher, a função discriminante foi calculada conforme expressa na Eq. (4).

$$\begin{aligned} D_m(\mathbf{X}) &= \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^T S^{-1} x \\ &\geq \frac{1}{2}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) S^{-1} \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) \end{aligned} \tag{4}$$

sendo $\bar{\mathbf{X}}_1 = [Q_1,\ K_1,\ N_1]$; $\bar{\mathbf{X}}_2 = [Q_2,\ K_2,\ N_2]$ os vetores associados a m-ésima ($m = 1, 2$) microrregião e S a matriz de covariância combinada

$$S = \frac{\left(n_1 - 1\right)S_1 + \left(n_2 - 1\right)S_2}{n_1 + n_2 - 2}.$$

Na suposição da heterogeneidade entre as covariâncias das variáveis, definidas em cada microrregião, adotou-se a função discriminante quadrática definida pela Eq. (5).

$$\begin{aligned} D_m(\mathbf{X}) &= -\frac{1}{2}x^T(S_1^{-1} - S_2^{-1})x + (\bar{\mathbf{X}}_1^T S_1^{-1} - \bar{\mathbf{X}}_2^T S_2^{-1})x - k \\ &\geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right), \end{aligned} \tag{5}$$

sendo,

$$k = \frac{1}{2}(\bar{\mathbf{X}}_1^T S_1^{-1}\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2^T S_2^{-1}\mu_2) + \frac{1}{2}\ln\left(\frac{|S_1|}{|S_2|}\right).$$

Onde S_m é a matriz de covariância da classe m, $\bar{\mathbf{X}}_m$ é o vetor de médias da classe m, p_m é a probabilidade a priori da classe m, para $m = 1, 2$ e $c(m|j)$ é a probabilidade condicional de pertencer à classe m dado que pertence a j, para $j = 1, 2$ e $j \neq m$.

Em ambas as situações, a matriz de confusão foi computada conforme o layout fornecido na Tabela 6.

Tabela 6: Matriz de Confusão

		Predito		
		Microrregião 1	Microrregião 2	
Obs.	Microrregião 1	a	b	n_1
	Microrregião 2	c	d	n_2
		$a + c$	$b + d$	N

Fonte: Os Autores (2023).

Além disso, as métricas descritas abaixo, Eq. (6), foram calculadas.

$$\begin{aligned} \text{Acurácia} &= \frac{a+d}{N} \\ \text{Precisão} &= \frac{a}{a+c} \\ \text{Sensibilidade} &= \frac{a}{a+b} \\ \text{Taxa Falso Positivo} &= \frac{c}{n_2} \\ \text{Especificidade} &= \frac{d}{c+d} \\ \text{Taxa Falso Negativo} &= \frac{b}{n_1}. \end{aligned} \tag{6}$$

Por fim, para a reprodução dos resultados, elaborou-se um script na linguagem R Core Team (2023) com auxilio do software RStudio Team (2022), seguindo os procedimentos detalhados nas Seções 2, 3.1 e 3.2.

4 Resultados e Discussões

4.1 Simulação das microrregiões considerando função de ligação logit

As Figs. 2 a 4, apresentam os valores de acurácia e precisão para os modelos esférico, exponencial e gaussiano, considerando dois cenários de tamanho de amostra: $n = 25$ e $n = 100$. Nesses gráficos, a representação dos resultados da análise discriminante linear está em azul, enquanto os da análise discriminante quadrática estão em rosa, com o eixo x denotando o índice de dependência espacial.

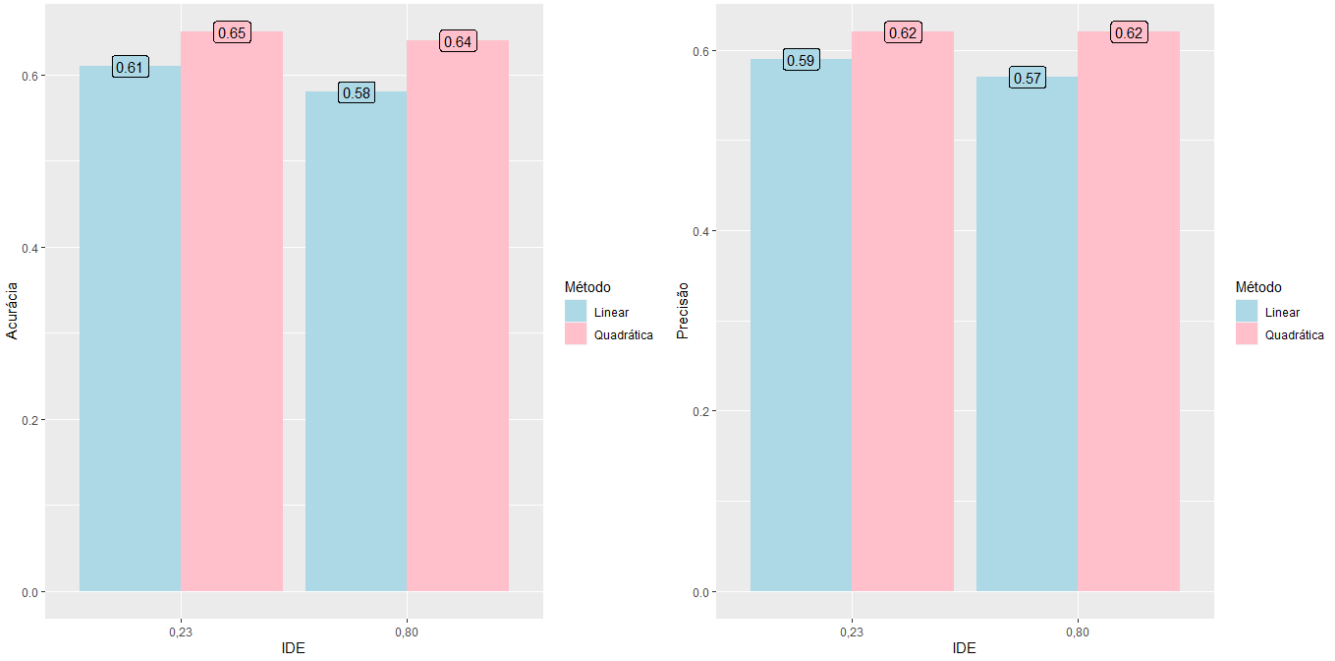
Ao analisar os vários modelos com $n = 25$, é evidente que, em situações de baixa dependência espacial, o uso do modelo exponencial resulta em taxas de acurácia superiores, contrastando com o modelo gaussiano, que apresenta as menores taxas, independentemente do tipo de análise discriminante adotada. No entanto, em cenários de alta dependência espacial, os modelos exponencial e gaussiano alternam-se nas melhores taxas de acurácia para ambos os tipos de análise.

Quanto à precisão, em contextos de baixa dependência, o desempenho do modelo gaussiano e exponencial varia conforme a análise discriminante utilizada. Contudo, em situações de alta dependência, o modelo exponencial é mais preciso.

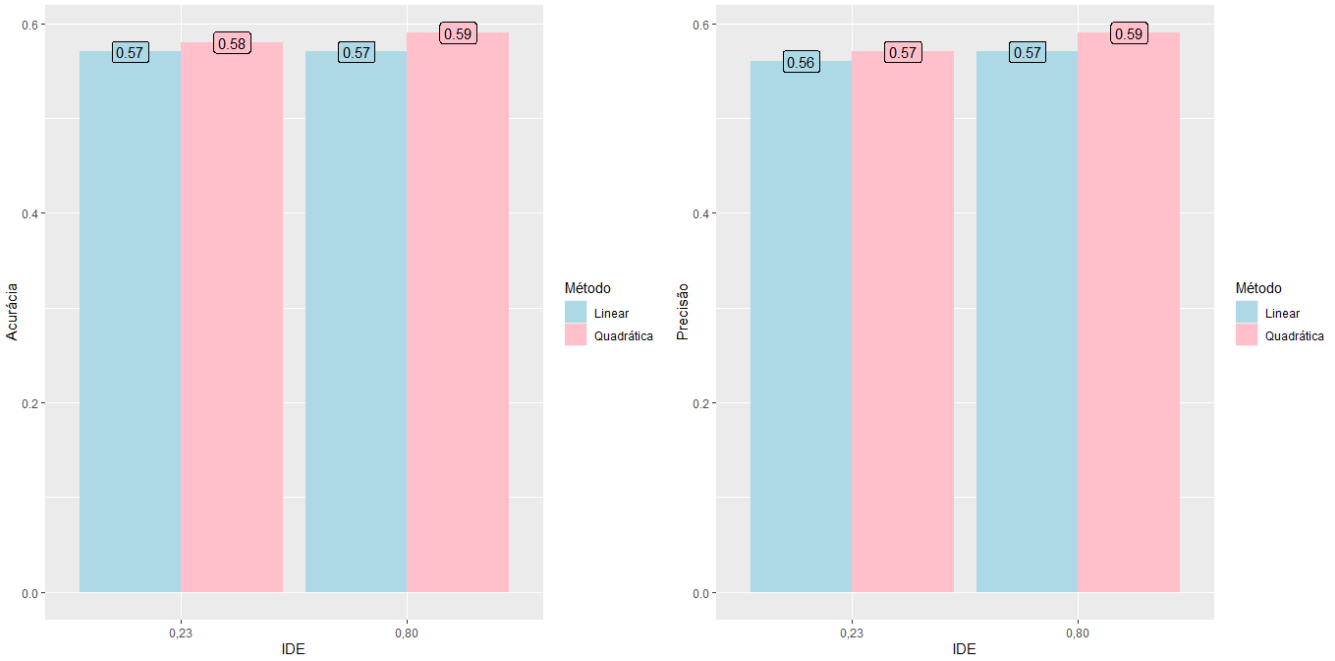
Quando $n = 100$, observa-se que as taxas de acurácia são melhores ao utilizar o modelo exponencial e menores para o modelo esférico. A precisão oscila de acordo com a dependência espacial adotada, sendo o modelo exponencial preferível para baixas dependências e o gaussiano para alta dependência espacial.

Na Tabela 7, são apresentados os índices de desempenho das simulações para $n = 25$ e 100 quando a dependência espacial é baixo. Observa-se um aumento nas taxas de falso negativo e especificidade com o aumento de n, enquanto a sensibilidade e a taxa de falso positivo diminuem. Os modelos exibem resultados semelhantes, sendo que a análise discriminante quadrática apresenta os melhores taxas para $n = 25$ e convergem com o aumento de n.

Figura 2: Gráficos de Acurácia e Precisão para o Modelo Esférico usando função de ligação logit.

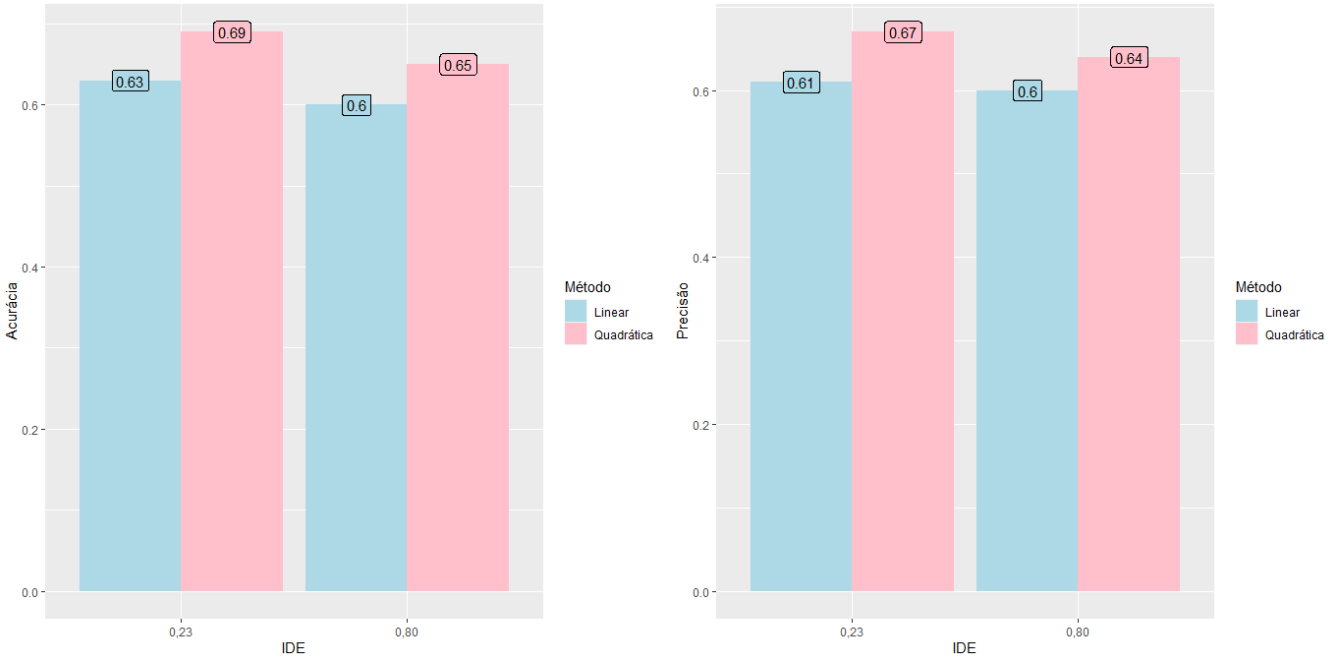


(a) n = 25

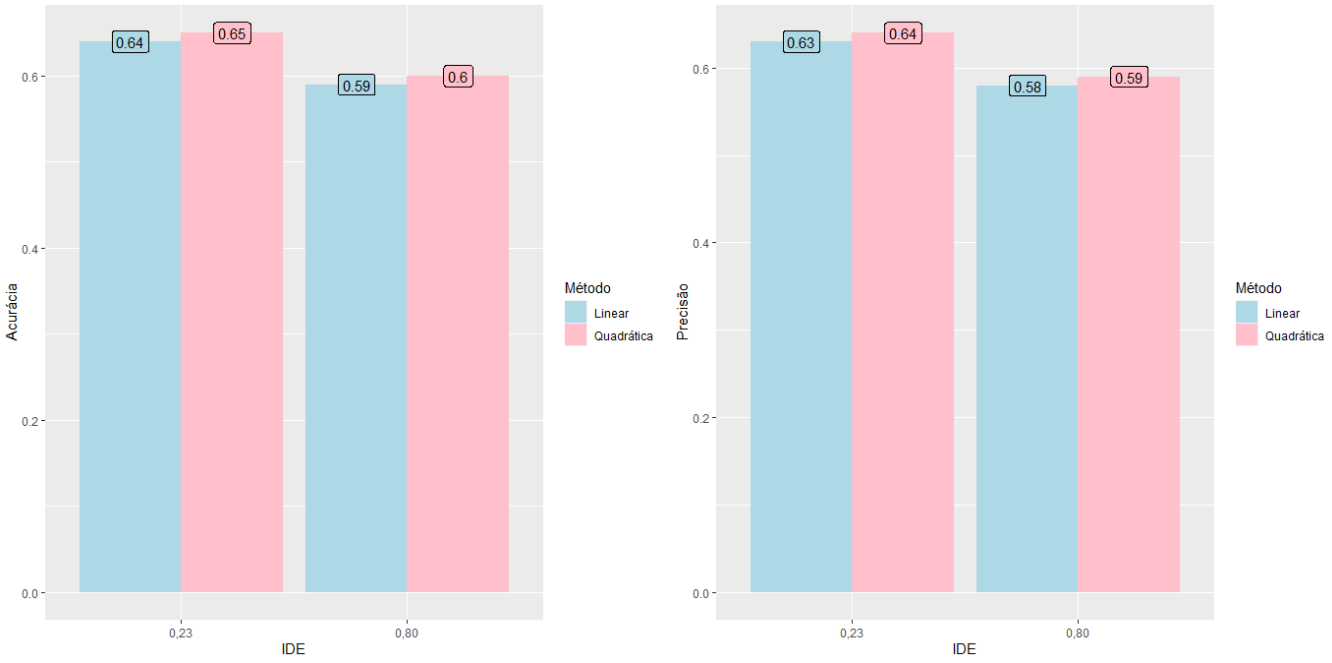


(b) n = 100
Fonte: Os autores (2024).

Figura 3: Gráficos de Acurácia e Precisão para o Modelo Exponencial usando função de ligação logit.



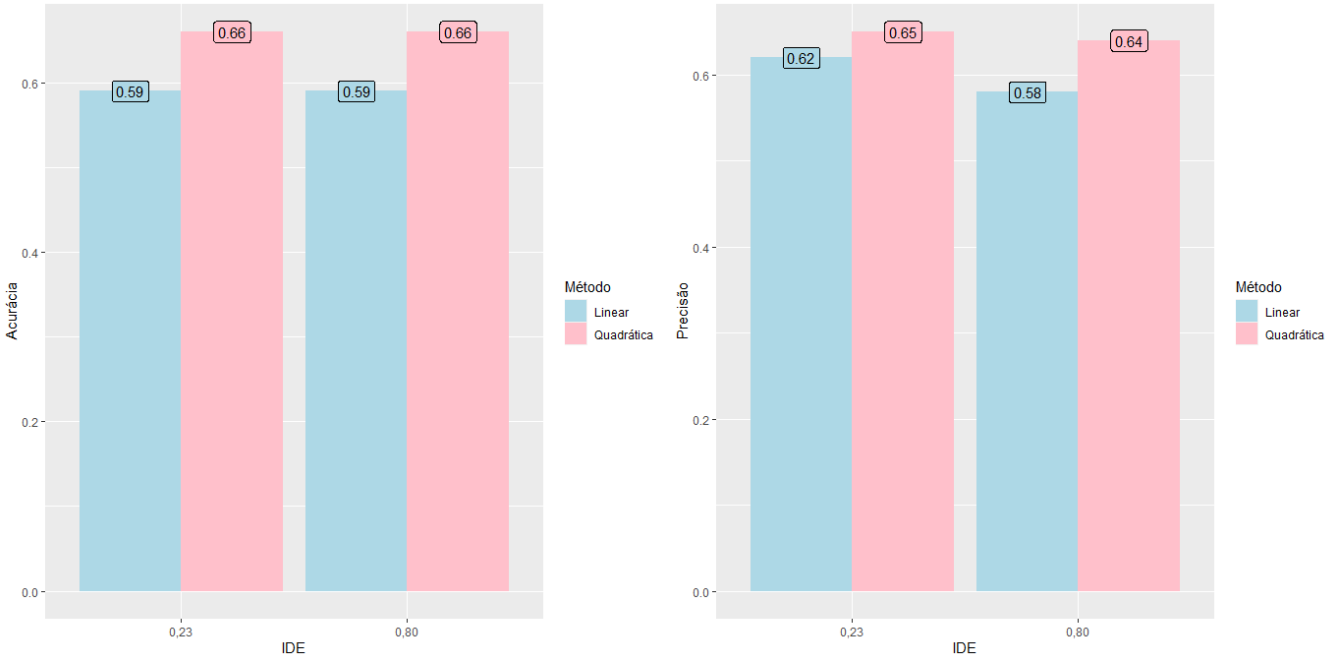
(a) n = 25



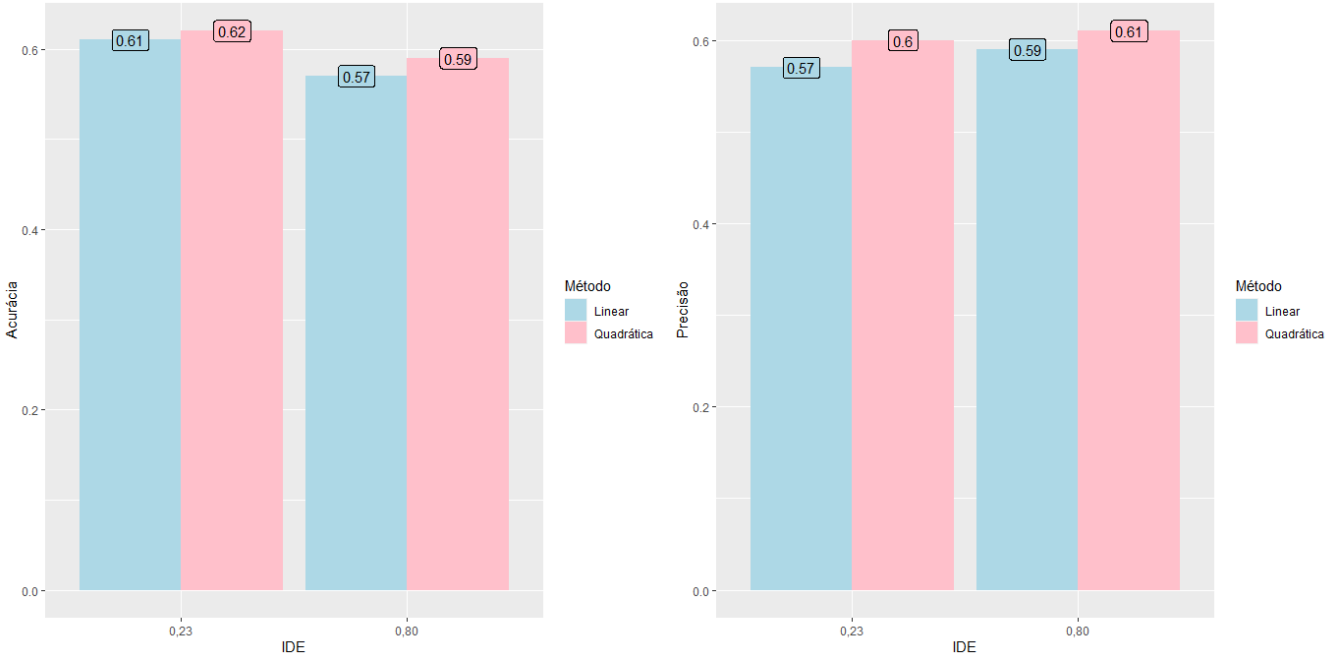
(b) n = 100

Fonte: Os autores (2024).

Figura 4: Gráficos de Acurácia e Precisão para o Modelo Gaussiano usando função de ligação logit.



(a) n = 25



(b) n = 100
Fonte: Os autores (2024).

Tabela 7: Tabela com os índices de Avaliação para Índice de Dependência Espacial fraco usando função de ligação logit.

Modelo		*n* = 25		*n* = 100	
		Linear	Quad	Linear	Quad
Esférico	TFP	0,46	0,45	0,45	0,43
	TFN	0,32	0,23	0,40	0,40
	Esp	0,54	0,55	0,55	0,57
	Sens	0,68	0,77	0,60	0,60
Gaussiano	TFP	0,46	0,45	0,45	0,43
	TFN	0,32	0,23	0,40	0,40
	Esp	0,54	0,55	0,55	0,57
	Sens	0,68	0,77	0,60	0,60
Exponencial	TFP	0,47	0,45	0,45	0,44
	TFN	0,32	0,23	0,40	0,40
	Esp	0,53	0,55	0,55	0,56
	Sens	0,68	0,77	0,60	0,60

Fonte: Os autores (2024).

O mesmo comportamento apresentado na fraca dependência espacial é observado na forte dependência (Tabela 8) para os índices de desempenho. No entanto, o aumento de *n* faz com que o modelo exponencial se diferencie dos demais, obtendo melhores resultados para a taxa de falso positivo e especificidade.

Tabela 8: Tabela com os índices de Avaliação para Índice de Dependência Espacial forte usando função de ligação logit.

Modelo		*n* = 25		*n* = 100	
		Linear	Quad	Linear	Quad
Esférico	TFP	0,49	0,41	0,46	0,40
	TFN	0,35	0,31	0,38	0,41
	Esp	0,51	0,59	0,54	0,60
	Sens	0,65	0,69	0,62	0,59
Gaussiano	TFP	0,49	0,41	0,46	0,40
	TFN	0,35	0,31	0,38	0,41
	Esp	0,51	0,59	0,54	0,60
	Sens	0,65	0,69	0,62	0,59
Exponencial	TFP	0,49	0,41	0,44	0,38
	TFN	0,35	0,31	0,38	0,41
	Esp	0,51	0,59	0,56	0,62
	Sens	0,65	0,69	0,62	0,59

Fonte: Os autores (2024).

4.2 Simulação das microrregiões considerando função de ligação complemento log-log

Os resultados da simulação, $n = 25$, utilizando a função de ligação complemento log-log mostram que os padrões comportamentais são semelhantes, sem grandes alterações em relação aos resultados anteriores obtidos.

Já para $n = 100$, destaca-se o modelo gaussiano fornecendo as melhores taxas de acurácia e precisão. Além disso, ao aumentar o valor de *n*, verifica-se uma tendência de convergência entre os resultados das análises discriminante linear e quadrática.

Os demais índices para os dados simulados com a função de ligação complemento log-log serão apresentados nas Tabelas 9 e 10, onde, de modo geral, ocorre uma leve queda na taxa de falso positivo e sensibilidade, e um incremento nas taxas de falso negativo e especificidade para os três modelos ao passar de $n = 25$ para 100. Além disso, para análise discriminante quadrática, a independência espacial não interfere nos resultados obtidos para $n = 100$.

Tabela 9: Tabela com os índices de Avaliação para Índice de Dependência Espacial fraco usando função de ligação complemento log-log.

Modelo		*n* = 25		*n* = 100	
		Linear	Quad	Linear	Quad
Esférico	TFP	0,52	0,53	0,50	0,44
	TFN	0,34	0,20	0,35	0,40
	Esp	0,48	0,47	0,50	0,56
	Sens	0,67	0,80	0,65	0,60
Gaussiano	TFP	0,52	0,53	0,50	0,44
	TFN	0,34	0,20	0,35	0,40
	Esp	0,48	0,47	0,50	0,56
	Sens	0,67	0,80	0,65	0,60
Exponencial	TFP	0,52	0,53	0,49	0,44
	TFN	0,34	0,20	0,37	0,40
	Esp	0,48	0,47	0,51	0,56
	Sens	0,67	0,80	0,63	0,60

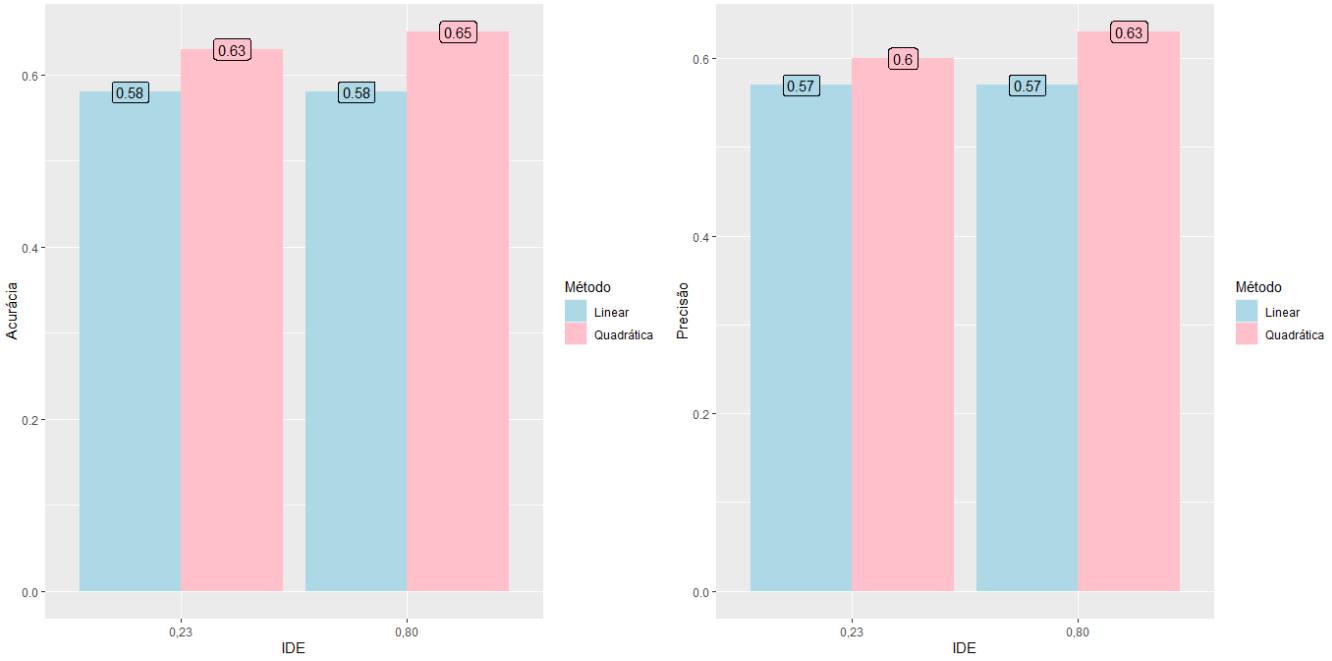
Fonte: Os autores (2024).

Tabela 10: Tabela com os índices de Avaliação para Índice de Dependência Espacial forte usando função de ligação complemento log-log.

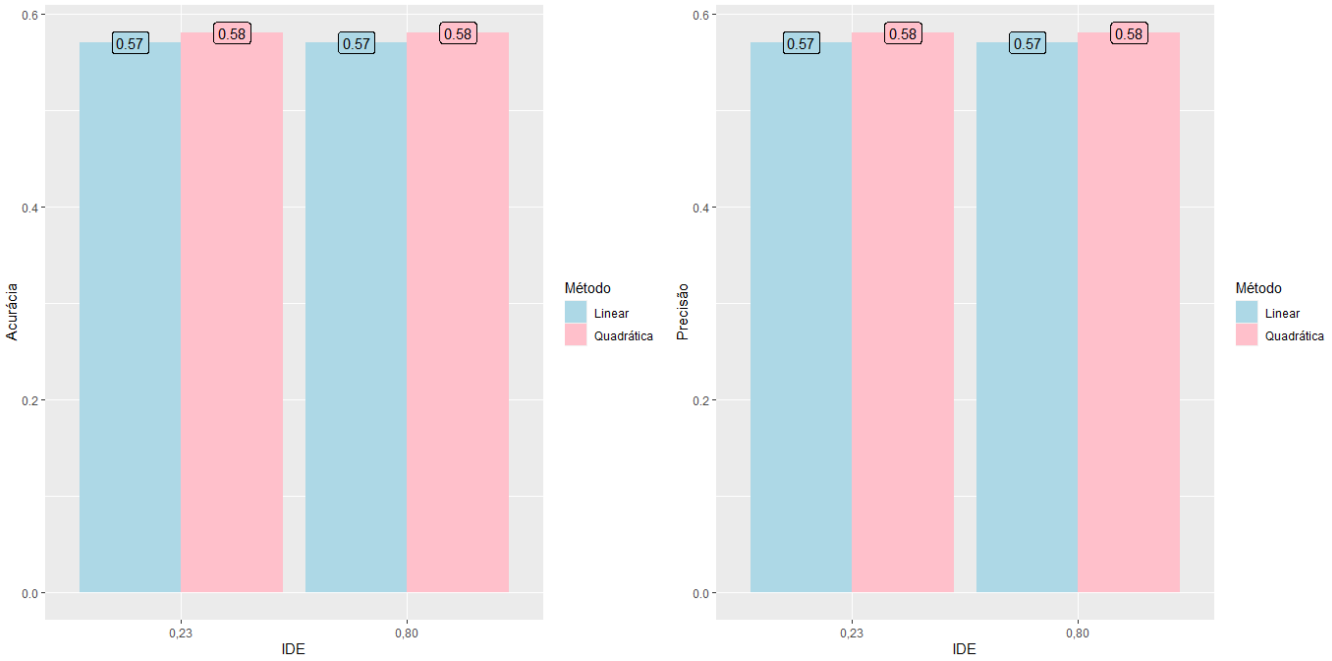
Modelo		*n* = 25		*n* = 100	
		Linear	Quad	Linear	Quad
Esférico	TFP	0,47	0,45	0,41	0,44
	TFN	0,38	0,26	0,45	0,40
	Esp	0,53	0,55	0,59	0,56
	Sens	0,62	0,74	0,55	0,60
Gaussiano	TFP	0,47	0,45	0,41	0,44
	TFN	0,38	0,26	0,45	0,40
	Esp	0,53	0,55	0,59	0,56
	Sens	0,62	0,74	0,55	0,60
Exponencial	TFP	0,47	0,45	0,41	0,44
	TFN	0,38	0,26	0,44	0,40
	Esp	0,53	0,55	0,59	0,56
	Sens	0,62	0,74	0,56	0,60

Fonte: Os autores (2024).

Figura 5: Gráficos de Acurácia e Precisão para o Modelo Esférico usando função de ligação complemento log-log.

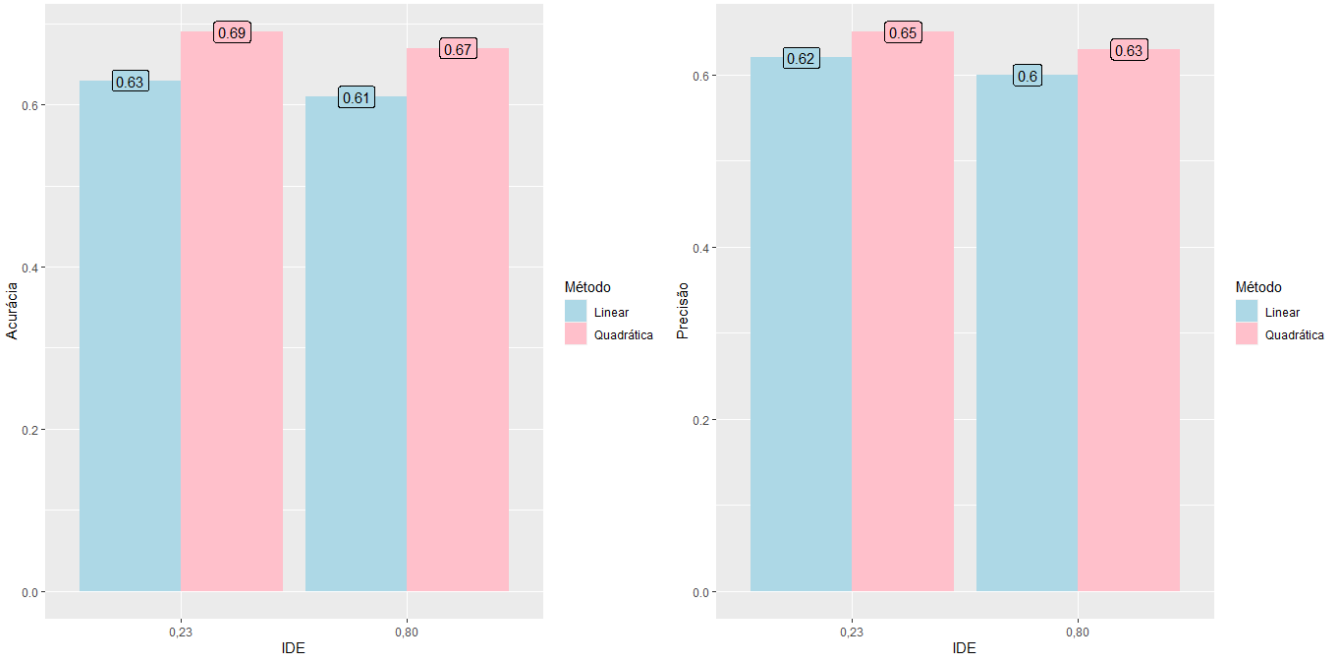


(a) n = 25

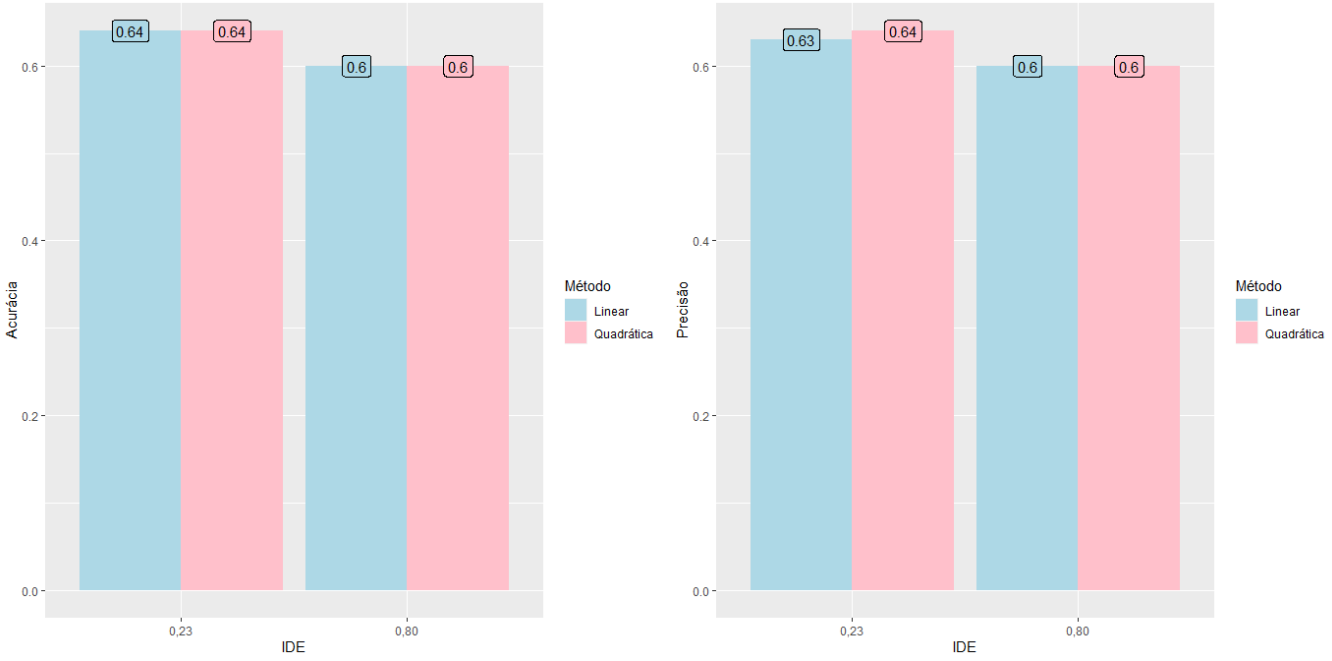


(b) n = 100
Fonte: As autores (2024).

Figura 6: Gráficos de Acurácia e Precisão para o Modelo Exponencial usando função de ligação complemento log-log.



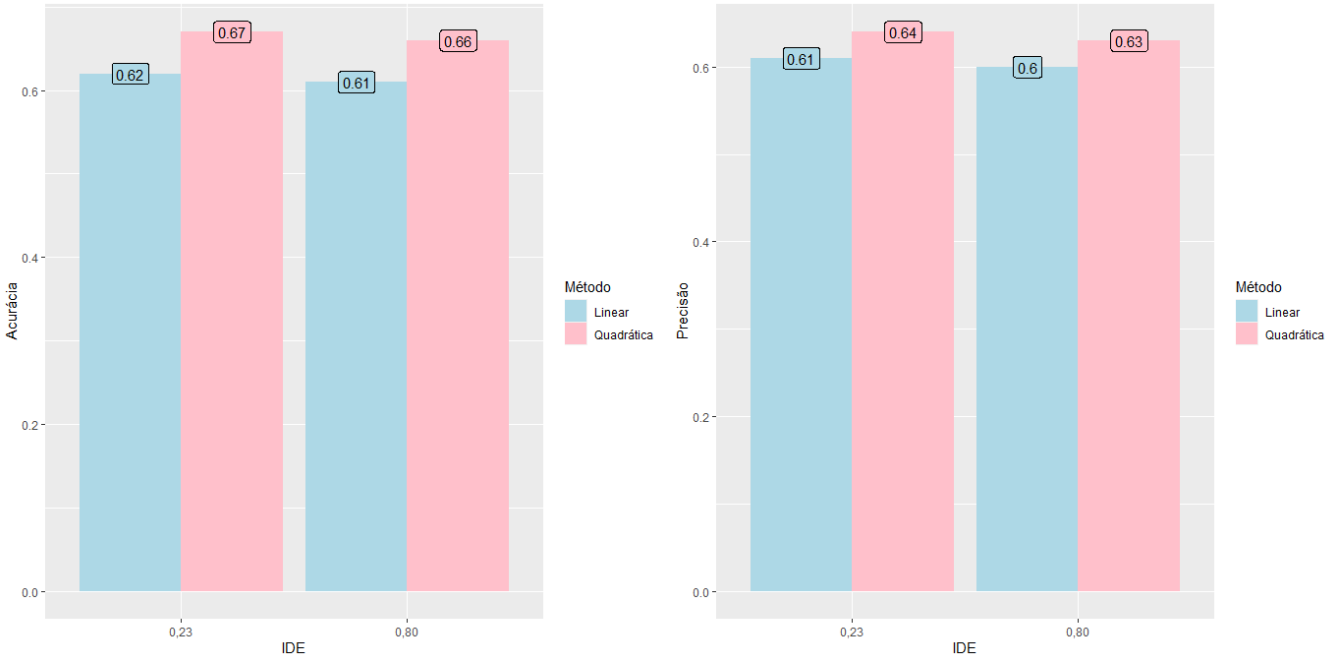
(a) n = 25



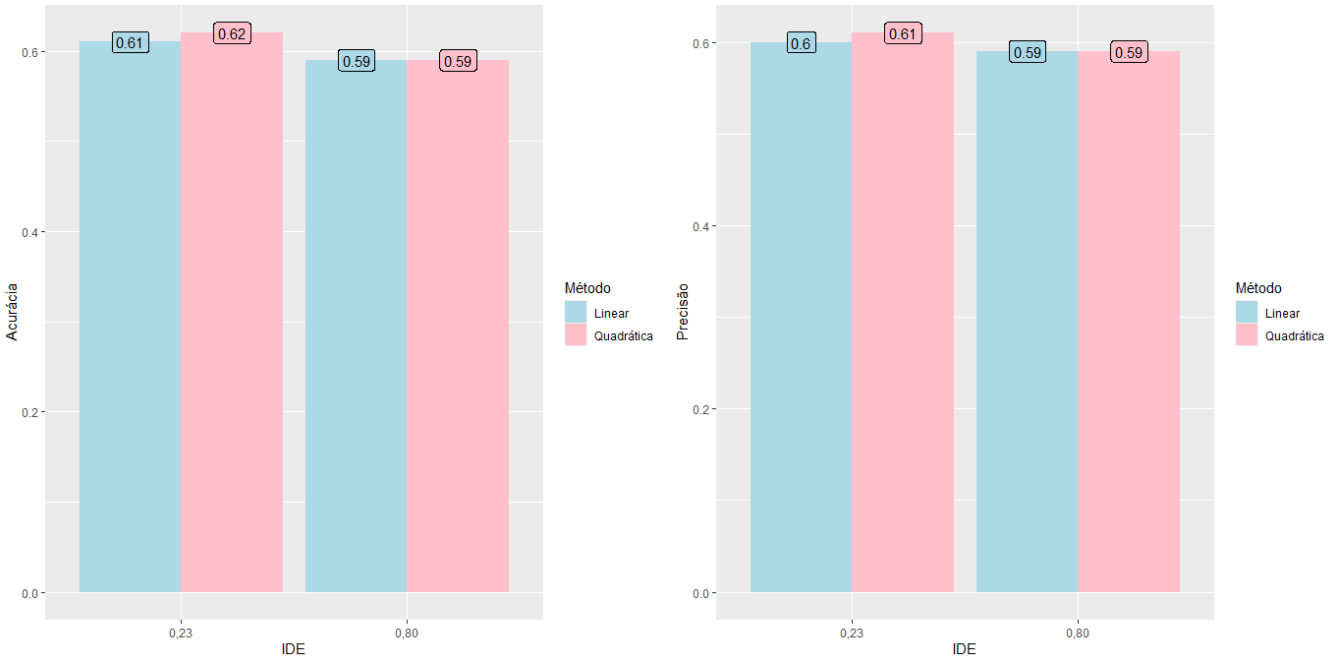
(b) n = 100

Fonte: Os autores (2024).

Figura 7: Gráficos de Acurácia e Precisão para o Modelo Gaussiano usando função de ligação complemento log-log.



(a) n = 25



(b) n = 100

Fonte: Os autores (2024).

5 Conclusão

Em conclusão, este artigo avaliou a eficácia dos modelos esférico, exponencial e gaussiano na análise de microrregiões simuladas, utilizando análises discriminantes linear e quadrática em dois cenários de dependência espacial e tamanhos de amostra (n = 25 e n = 100). Nestas conjecturas, a metodologia computacional proposta respeitou os critérios da ABIC que diferencia a qualidade de cafés, dessa forma, sugere-se que em uma analise preliminar, antes de proceder com as análises de dados reais o procedimento computacional abordado neste trabalho, poderá ser utilizado, como instrumento de análise exploratória em relação a análise da qualidade.

Os resultados mostraram que, para amostras pequenas, o modelo exponencial teve o melhor desempenho em termos de acurácia em contextos de baixa dependência espacial, enquanto o modelo gaussiano se destacou em alta dependência espacial. Com o aumento do tamanho da amostra para n = 100, o modelo exponencial manteve sua superioridade em baixa dependência, e o modelo gaussiano demonstrou melhores resultados em alta dependência.

A comparação entre as funções de ligação logit e complemento log-log revelou que a função logit foi mais eficaz na maioria dos cenários analisados. Embora a função complemento log-log apresentasse desempenho semelhante para amostras pequenas, o modelo gaussiano mostrou melhores taxas de acurácia e precisão para amostras maiores. Esses achados destacam a importância de adaptar a escolha do modelo e da função de ligação com base na dependência espacial dos dados e no tamanho da amostra.

Além disso, a análise discriminante quadrática provou ser consistentemente superior à linear, especialmente para amostras menores. Os resultados deste estudo oferecem diretrizes úteis para a seleção de métodos apropriados para análise espacial em microrregiões e ressaltam a necessidade de uma abordagem personalizada para diferentes contextos espaciais.

Limitações e Próximos Passos: Apesar das contribuições do estudo, a análise foi realizada com dados simulados. Como próximos passos, pretende-se aplicar os métodos a dados reais de duas regiões distintas para validar sua eficácia em contextos práticos.

Agradecimentos

Projeto PQ - 304939/2021-8 financiado pelo CNPq;
Fonte de Financiamento: CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Referências

ABIC (2020). Programa de qualidade do café: Norma da qualidade recomendável e boas práticas de fabricação de cafés torrados em grão e cafés torrados e moídos, Rio de Janeiro: ABIC. Disponível em http://abic.com.br/.

Bouveyron, C., Girard, S. and Schimid, C. (2007). High dimensional discriminant analysis, *Communications in Statistics-Theory and Methods* **36**(14): 2607–2623. Disponível em https://www.researchgate.net/publication/228668717_High_Dimensional_Discriminant_Analysis.

Cambardella, C., Moorman, T., Novak, J., Parkin, T. and Karlen, D. (1994). Field-scale variability of soil properties in central iowa soils, *Soil Science Society of America Journal* **58**(5): 1501–1511. https://doi.org/10.2136/sssaj1994.03615995005800050033x.

de Oliveira, L. K., Resende, M., Borém, F. M. and Cirillo, M. A. (2023). Feasibility of the expectation-maximization algorithm for assessing individuals with different sensory perceptions in discrimination of specialty coffees, *Acta Scientiarum. Technology* **45**(1): e60184. https://doi.org/10.4025/actascitechnol.v45i1.60184.

Faria, P. N., Dias, C. T. S., Pinheiro, J. B., Cirillo, M. A., Araujo, L. B. and Araujo, M. C. (2016). Ammi methodology in soybean: Cluster analysis with bootstrap resampling in genetic divergence and stability, *REVISTA CERES* **63**: 461–468. https://doi.org/10.1590/0034-737X201663040005.

Hair Jr., J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2019). *Multivariate Data Analysis*, 8th edition edn, Pearson.

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*, Applied Multivariate Statistical Analysis, Pearson Prentice Hall. https://books.google.com.br/books?id=gFWcQgAACAAJ.

Manoel, I. d. S., Resende, M., Sousa, P. H. A., Rosa, S. D. V. F. d. and Cirillo, M. A. (2024). Simulation of robust adaptive regression multi-level models for quality analysis of special coffees in cold storage, *Acta Scientiarum. Technology* **46**(1): e59135. Disponível em https://periodicos.uem.br/ojs/index.php/ActaSciTechnol/issue/view/2040.

Menezes, F. S. d., Liska, G. R., Cirillo, M. A. and Vivanco, M. J. (2016). Data classification with binary response through the boosting algorithm and logistic regression, *Expert Systems with Applications* **65**: 169–179. https://doi.org/10.1016/j.eswa.2016.08.014.

Oliveira, L. M. d., Menezes, F. S. d., Cirillo, M. A., Saúde, A. V., Borém, F. M. and Liska, G. R. (2019). Machine learning techniques in multiclass problems with application in sensorial analysis, *Concurrency and Computation: Practice and Experience* . https://doi.org/10.1002/cpe.5579.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Disponível em https://www.R-project.org/.

Rausch, J. R. and Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality, *Behavior Research Methods* **41**(1): 85–98. https://doi.org/10.3758/BRM.41.1.85.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*, RStudio, PBC. Disponível em https://www.rstudio.com/.