UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Larissa Carvalho Solino Silva

Fatores Associados à Presença de Triatomíneos em Comunidades Rurais de Goiás: Uma Abordagem com Modelagem Estatística

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Larissa Carvalho Solino Silva

Título do trabalho: Fatores Associados à Presença de Triatomíneos em Comunidades Rurais de Goiás: Uma Abordagem com Modelagem Estatística

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Luis Rodrigo Fernandes Baumann**, **Professor do Magistério Superior**, em 16/12/2025, às 07:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Larissa Carvalho Solino Silva**, **Discente**, em 18/12/2025, às 09:18, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Larissa Carvalho Solino Silva

Fatores Associados à Presença de Triatomíneos em Comunidades Rurais de Goiás: Uma Abordagem com Modelagem Estatística

Trabalho de Conclusão de Curso apresentado ao Instituto de Matemática e Estatística da Universidade Federal de Goiás como requisito parcial para obtenção do grau de Bacharel em Estatística.
Orientador: Dr. Luis Rodrigo Fernandes Baumann

Goiânia
2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Silva, Larissa Carvalho Solino
Fatores associados à presença de triatomíneos em comunidades rurais de Goiás: uma abordagem com modelagem estatística [manuscrito] / Larissa Carvalho Solino Silva. - 2025.
63 f.

Orientador: Prof. Dr. Luis Rodrigo Fernandes Baumann .
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Apêndice.
Inclui siglas, gráfico, tabelas, lista de figuras, lista de tabelas.

1. Doença de Chagas . 2. Triatomíneos. 3. Regressão logística. 4. Determinantes sociais. I. Baumann , Luis Rodrigo Fernandes, orient. II. Título.

CDU 519.22

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e seis dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Fatores Associados à Presença de Triatomíneos em Comunidades Rurais de Goiás: Uma Abordagem com Modelagem Estatística”, de autoria de Larissa Carvalho Solino Silva, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Prof. Dr. Luis Rodrigo Fernandes Baumann com a participação dos demais membros da Banca Examinadora: David Henriques da Matta (IME/UFG) e Marcílio Ramos Pereira Cardial (IME/UFG). Após a apresentação, a banca examinadora realizou a arguição da estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 7,7 , tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Luis Rodrigo Fernandes Baumann**, **Professor do Magistério Superior**, em 16/12/2025, às 07:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 16/12/2025, às 09:13, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Marcilio Ramos Pereira Cardial**, **Professor do Magistério Superior**, em 16/12/2025, às 09:38, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5801921** e o código CRC **0204BF6E**.

Referência: Processo nº 23070.060277/2025-20 SEI nº 5801921

Agradecimentos

Gostaria de expressar minha gratidão ao meu orientador, Dr. Luis Rodrigo Fernandes Baumann, pela orientação dedicada, paciência e pelos valiosos ensinamentos ao longo de toda a elaboração deste trabalho. Agradeço também à Universidade Federal de Goiás e ao Instituto de Matemática e Estatística pelo apoio institucional.

Meus agradecimentos se estendem à equipe do projeto SanRural, pela disponibilização dos dados e pelo suporte.

Finalmente, agradeço à minha família e amigos pelo incentivo constante e pelo apoio emocional durante toda esta jornada acadêmica, em especial à minha mãe, Marenilde, e à minha tia, Marlene, que sempre estiveram ao meu lado, apoiando-me com dedicação e carinho.

Resumo

A doença de Chagas permanece como um desafio relevante no meio rural brasileiro, sobretudo em comunidades tradicionais. Este estudo investigou fatores sociais e habitacionais associados à presença de triatomíneos em domicílios de Goiás, com base nos dados do projeto SanRural, abrangendo quilombolas, assentamentos e ribeirinhos. Foram analisadas variáveis como tipo de comunidade, material das paredes, estrutura etária dos moradores, uso de caixa d'água ou amianto e criação de animais domésticos. A abordagem estatística incluiu regressão logística, modelos aditivos generalizados - regressão logística (GAM-LR). A regressão logística permitiu identificar associações diretas entre os fatores avaliados e a infestação; o GAM-LR evidenciou efeitos não lineares relacionados à idade do responsável pelo domicílio e à razão moradores/quartos. Os resultados revelaram maior proporção de casas infestadas entre comunidades quilombolas (14,3%), além da influência da presença de crianças pequenas e de histórico positivo para doença de Chagas entre os residentes. Os modelos apresentaram bom desempenho (acurácia de 86,1% e especificidade de 97,3%), reforçando que o risco de infestação está intimamente relacionado às condições sociais e estruturais das moradias.

Palavras-chave: doença de Chagas; triatomíneos; regressão logística; determinantes sociais.

Abstract

Chagas disease remains a significant public health challenge in rural areas of Brazil, particularly among traditional communities. This study investigated social and housing factors associated with the presence of triatomine insects in households in the state of Goiás, using data from the SanRural project, which includes quilombola, settlement, and riverside communities. Variables analyzed comprised type of community, wall construction materials, household age structure, use of water tanks or asbestos roofing, and the presence of domestic animals. The statistical approach included logistic regression and generalized additive models with a logistic link (GAM-LR). Logistic regression identified direct associations between the evaluated factors and household infestation, while GAM-LR revealed non-linear effects related to the age of the household head and the residents-to-rooms ratio. The results showed a higher proportion of infested houses among quilombola communities (14.3%), as well as the influence of the presence of young children and a positive history of Chagas disease among residents. The models demonstrated good performance, with an accuracy of 86.1% and a specificity of 97.3%, reinforcing that infestation risk is closely linked to the social and structural conditions of dwellings.

Keywords: Chagas disease; triatomines; logistic regression; social determinants.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

DC	Doença de Chagas
SINAN	Sistema de Informação de Agravos de Notificação
DATASUS	Departamento de Informática do Sistema Único de Saúde
GLM	Generalized Linear Model (Modelo Linear Generalizado)
LR	(Logistic Regression) Regressão Logística
GAM	Generalized Additive Model (Modelo Aditivo Generalizado)
UFG	Universidade Federal de Goiás
FUNASA	Fundação Nacional de Saúde

Sumário

1 INTRODUÇÃO

A doença de Chagas (DC) é uma das principais endemias parasitárias da América Latina e continua sendo um problema relevante de saúde pública no Brasil (WHO, 2025). Descrita pela primeira vez em 1909 pelo médico Carlos Chagas, a doença foi identificada de maneira pioneira, incluindo simultaneamente o agente etiológico, Trypanosoma cruzi, o vetor transmissor, insetos da subfamília Triatominae ou Triatomíneos (popularmente conhecidos como barbeiros), e a forma de transmissão vetorial, que ocorre por meio das fezes do inseto depositadas após o repasto sanguíneo em condições precárias de moradia, como casas de pau a pique ou de adobe típicas de áreas rurais (KROPF, 2009; DOS SANTOS et al., 2020).

A principal via de transmissão da DC permanece sendo a vetorial. O triatomíneo, ao se alimentar de sangue, elimina fezes infectadas nas proximidades do local da picada, permitindo a entrada do parasita na corrente sanguínea do hospedeiro (JURBERG et al., 2014). Embora outras formas de transmissão congênita, transfusional, por transplante de órgãos e acidentes laboratoriais tenham relevância, a transmissão vetorial mantém papel central, sobretudo em comunidades rurais e populações tradicionais que vivem em maior vulnerabilidade (DALE; PASCHOALETTO; COSTA, 2019).

A presença do vetor está fortemente associada a fatores socioeconômicos, ambientais e estruturais. Historicamente, a doença de Chagas tem sido vinculada à pobreza, marginalização social e condições habitacionais precárias (CHAO; LEONE VIGLIANO, 2022). Moradias construídas com adobe, madeira ou materiais vegetais, comuns em comunidades quilombolas, ribeirinhas e assentamentos, oferecem abrigos ideais para os triatomíneos (DORN; MONROY; STEVENS, 2022). Adicionalmente, práticas como armazenamento inadequado de materiais, criação de animais no peridomicílio e proximidade de áreas de mata aumentam a exposição ao vetor (BATISTA; LIMA, 2009).

No estado de Goiás, historicamente considerado endêmico, o Sistema de Informação de Agravos de Notificação (SINAN) registrou casos agudos confirmados de doença de Chagas em dez municípios entre 2010 e 2019 (SINAN, 2020). Nesse contexto, o projeto SanRural(Saneamento e Saúde Ambiental Rural), realizado pela Universidade Federal de Goiás (UFG) em parceria com a Fundação Nacional de Saúde (FUNASA), analisou comunidades tradicionais em 45 municípios do estado, identificando indicadores de risco para a ocorrência de triatomíneos e evidenciando a importância da vigilância contínua. Foi a partir desse projeto que se originou a base de dados utilizada para o presente trabalho.

Nesse contexto, compreender os determinantes ambientais, estruturais e socioeconômicos associados à presença domiciliar de triatomíneos é fundamental para subsidiar

políticas públicas de prevenção e controle da doença de Chagas. O emprego de técnicas de modelagem estatística, como a regressão logística baseada em Modelos Lineares Generalizados (GLM-LR) e os Modelos Aditivos Generalizados com função de ligação logito (GAM-LR), Possibilita quantificar a força de associação entre variáveis explicativas e a presença do vetor, identificar relações lineares e não lineares e estimar, com maior precisão, os domicílios de maior risco.

Além de fortalecer a compreensão científica da dinâmica da transmissão vetorial, tais análises orientam intervenções mais direcionadas e eficientes, otimizando recursos em ações de vigilância entomológica, saneamento e educação em saúde. Assim, contribuem para estratégias de prevenção da doença de Chagas mais eficazes em comunidades rurais vulneráveis.

Portanto, este trabalho está estruturado em quatro capítulos principais, além de referências e apêndice. O Capítulo 1 apresenta a introdução. O Capítulo 2 aborda o referencial teórico e a metodologia, contemplando a biologia dos triatomíneos, determinantes sociais e ambientais, a epidemiologia da doença de Chagas em Goiás e as técnicas de modelagem empregadas (GLM-LR e GAM-LR). O Capítulo 3 reúne os resultados, incluindo as análises descritivas, a aplicação dos três modelos e a comparação de seu desempenho. O Capítulo 4 traz a conclusão, destacando os principais achados, as contribuições do estudo, suas implicações para políticas públicas e as limitações. Por fim, o trabalho inclui as referências utilizadas e, no apêndice, o script completo em R com as etapas de análise.

2 Referencial Teórico e Metodologia

2.1 Triatomíneo: Biologia e Papel na Transmissão da Doença de Chagas

Os triatomíneos, popularmente conhecidos como barbeiros, constituem os principais vetores da doença de Chagas, enfermidade causada pelo protozoário *Trypanosoma cruzi*. Pertencem à subfamília Triatominae, da ordem Hemiptera, e apresentam ampla distribuição geográfica na América Latina, incluindo Brasil, México e diversos países da América Central e do Sul (JURBERG, 2014; DORN, 2022). A densidade populacional desses insetos é geralmente maior em áreas rurais e periféricas, refletindo fatores como precariedade das construções habitacionais, presença de animais domésticos e proximidade de áreas de vegetação silvestre.

O ciclo de vida dos triatomíneos é composto por três fases: ovos, cinco estágios ninfais e a fase adulta. Todas as fases, após a eclosão, apresentam hábito hematófago, o que contribui para a manutenção eficiente do *Trypanosoma cruzi* em ambientes domiciliares e peridomiciliares (DIAS, 2006). Esses insetos destacam-se ainda por sua alta fecundidade, longevidade relativa e capacidade de adaptação a diferentes tipos de abrigos, desde tocas de animais silvestres até habitações humanas improvisadas (MONCAYO, 2003).

A transmissão da doença de Chagas ocorre principalmente quando as fezes contaminadas do triatomíneo entram em contato com mucosas ou feridas na pele durante ou após a picada. Tal mecanismo evidencia a relevância da infestação domiciliar como risco epidemiológico central. Estudos apontam que a abundância de triatomíneos está fortemente associada à presença de animais domésticos, à proximidade de áreas silvestres e ao acúmulo de materiais orgânicos ou entulho no peridomicílio (BATISTA, 2009; COURA, 2010).

Portanto, compreender a biologia e o comportamento dos triatomíneos é fundamental para a formulação de estratégias de prevenção e controle da doença de Chagas, especialmente em regiões onde o ciclo de transmissão ainda encontra condições favoráveis.

2.2 Determinantes sociais e ambientais

A presença de triatomíneos e a transmissão da doença de Chagas são influenciadas por fatores sociais e ambientais, tais como:

- Condições habitacionais como paredes de adobe, pau a pique, palha ou outros ma-

teriais vegetais aumentam a probabilidade de infestação.

- Saneamento básico precário como ausência de água encanada, esgoto e manejo inadequado de lixo favorecem a manutenção do vetor.
- Fatores demográficos como idade do responsável pelo domicílio, densidade de moradores, composição familiar e ocupação dos quartos influenciam o risco de exposição.
- Criação de animais domésticos como cães, gatos, galinhas e roedores próximos às residências criam abrigo e alimento para o vetor (DIAS, 2006; COURA, 2010).

No estado de Goiás, o projeto SanRural(2018-2019) investigou comunidades rurais tradicionais como assentamentos, quilombolas e ribeirinhas e demonstrou que a interação entre determinantes sociais e ambientais favorece a manutenção do ciclo domiciliar do vetor da doença de Chagas. Os achados de 2018-2019 indicaram que as comunidades quilombolas apresentaram maior proporção de domicílios infestados em comparação aos assentamentos, possivelmente em função da precariedade habitacional, da composição familiar ampliada e da maior densidade populacional.

2.3 Epidemiologia da doença de Chagas em Goiás e comparação com outros estados

Historicamente, Goiás apresenta níveis moderados de endemia para a doença de Chagas, com registros de casos agudos e maior prevalência em áreas rurais. Dados do Sistema de Informação de Agravos de Notificação (SINAN) e do DATASUS indicam distribuição heterogênea dos casos, concentrando-se em municípios com maiores áreas de cultivo agrícola e proximidade com mata nativa.

Comparando com outros estados do Centro-Oeste e Sudeste, como Minas Gerais e Mato Grosso, Goiás apresenta menor densidade de triatomíneos domiciliares, embora certas regiões, principalmente assentamentos rurais e comunidades quilombolas, mantenham risco elevado (COURA, 2010; SILVEIRA, 2015).

2.4 Modelagem estatística

Neste trabalho, empregamos dois métodos para modelar a presença de triatomíneos nos domicílios: A regressão logística, modelada como um caso particular de Modelo Linear Generalizado (GLM-LR) e modelos aditivos generalizados - regressão logística (GAM-LR). A seguir, apresentamos a formulação matemática de cada método, bem como os procedimentos para predição, classificação cutoff via MCC e validação LOOCV.

2.4.1 Modelo Linear Generalizado - Regressão Logística (GLM-LR)

Considere k variáveis explicativas independentes $X_1, \ldots, X_k$ e uma variável resposta Y. Um modelo linear generalizado (GLM) é dado por

$$Y_i \overset{\text{ind}}{\sim} \mathcal{FE}(\mu_i, \phi), \qquad \text{tal que} \qquad g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij},$$

onde $\mathcal{FE}(\mu_i, \phi)$ denota uma distribuição da família exponencial com média μ_i e parâmetro de escala ϕ; g é uma função de ligação; k é a quantidade de variáveis explicativas e β_i são coeficientes do modelo linear (Wood, S. N., 2017).

A regressão logística é um caso particular de GLM, a regressão logística é um caso particular do GLM em que a variável resposta é binária. No contexto deste trabalho, considera-se a variáveis resposta

$$Y_i = \begin{cases} 1, & \text{se o domicílio } i \text{ apresenta triatomíneos,} \\ 0, & \text{caso contrário.} \end{cases}$$

com função de ligação logito

$$g(\mu_i) = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}.$$

Nesse caso, $\mathcal{FE}(\mu_i, \phi)$ corresponde à distribuição Bernoulli com média $\mu_i = \pi_i$ e parâmetro de escala $\phi = 1$, caracterizando o modelo de regressão logística como um GLM com resposta binária e ligação logito.

Para observações independentes $Y_i \sim \text{Bernoulli}(\pi_i)$, a função de verossimilhança e o logaritmo da verossimilhança são

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)}, \qquad \ell(\beta) = \sum_{i=1}^{n} \left[Y_i \log(\pi_i) + (1-Y_i)\log(1-\pi_i)\right].$$

A estimativa de máxima verossimilhança $\hat{\beta}$ é obtida resolvendo $\partial\ell/\partial\beta = 0$, normalmente por meio do algoritmo Iteratively Reweighted Least Squares (IRLS) (DOBSON, 2018).

O erro padrão das estimativas dos coeficientes é obtido a partir da matriz de variâncias e covariâncias de $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^\top$, tal que:

$$\text{SE}(\hat{\beta}_j) = \sqrt{\left[\text{Var}(\hat{\boldsymbol{\beta}})\right]_{jj}}, \qquad j = 0, 1, \ldots, k,$$

onde $\left[\text{Var}(\hat{\boldsymbol{\beta}})\right]_{jj}$ denota o j-ésimo elemento da diagonal da matriz $\text{Var}\,\hat{\boldsymbol{\beta}})$. No modelo logístico, essa matriz é aproximada pelo inverso da matriz de informação observada:

$$\text{Var}(\hat{\boldsymbol{\beta}}) \approx \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1},$$

em que $\mathcal{I}(\boldsymbol{\beta})$ é a matriz de informação observada, dada por

$$\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top} = X^\top W X, \qquad w_{ii} = \pi_i(1-\pi_i),$$

onde X é a matriz de delineamento de dimensão $n \times (k+1)$, $\pi_i = P(Y_i = 1 \mid \mathbf{x}_i)$ é a probabilidade de sucesso para a observação i e W é uma matriz diagonal de pesos definida por

$$W = \operatorname{diag}\Big(\pi_1(1-\pi_1),\, \pi_2(1-\pi_2),\, \ldots,\, \pi_n(1-\pi_n)\Big).$$

Cada elemento w_{ii} representa a variância da distribuição Bernoulli para a observação i, refletindo o peso da contribuição dessa observação na estimação por máxima verossimilhança.

O intervalo de confiança de 95% para o coeficiente β_j é obtido por

$$IC_{95\%}(\beta_j) = \left[\hat{\beta}_j - z_{0{,}975}\, SE(\hat{\beta}_j),\ \hat{\beta}_j + z_{0{,}975}\, SE(\hat{\beta}_j)\right].$$

Uma maneira de avaliar a significância estatística dos coeficientes é por meio do *Teste de Wald*. Seja $\hat{\beta}_j$ o estimador de máxima verossimilhança do coeficiente associado à variável explicativa X_j, e $SE(\hat{\beta}_j)$ o seu erro-padrão. A estatística do teste é definido por

$$z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.$$

Sob a hipótese nula ($H_0 : \beta_j = 0$) e a estatística de Wald segue assintoticamente uma distribuição normal padrão, tal que $z_j \sim \mathcal{N}(0,1)$. O valor-p do teste é dado por:

$$p\text{-valor} = 2\left[1 - \Phi(|z_j|)\right].$$

A interpretação dos coeficientes é realizada por meio da *Razão de Chances*. As *odds* (razão de chances) de sucesso são definidas por

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i},$$

onde temos

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij},$$

isto é, o log das odds é modelado como uma função linear das covariáveis. Assim, cada coeficiente β_j pode ser interpretado como a variação no *log* das *odds* associada a um incremento unitário em X_{ij}, mantendo as demais covariáveis constantes. Tomando a exponencial, obtém-se o *odds ratio* (razão de chances) associado ao coeficiente β_j:

$$\text{OR}_j = \exp(\beta_j).$$

Para uma covariável contínua x_j, $\exp(\beta_j)$ representa o fator pelo qual as odds de sucesso são multiplicadas quando x_j aumenta em uma unidade, mantendo as demais variáveis fixas. Sem perca de generalidade, para uma covariável dicotômica $X_j \in \{0, 1\}$, $\exp(\beta_j)$corresponde ao *odds ratio* entre o grupo $X_j = 1$ e o grupo de referência $X_j = 0$, Matendo fixas as demais covariáveis do modelo.

A interpretação do *odds ratio* associado ao preditor X_j é dada por:

- $OR_j > 1$: um aumento de uma unidade em X_j (mantidas as demais covariáveis constantes) *eleva* as chances de ocorrência do evento;
- $OR_j < 1$: um aumento de uma unidade em X_j *reduz* as chances de ocorrência do evento;
- $OR_j = 1$: variações em X_j não alteram as chances de ocorrência do evento, indicando ausência de efeito associado a esse preditor.

O intervalo de confiança correspondente para o *odds ratio* é obtido exponenciando os limites:

$$\Big(\exp(\hat{\beta}_j - z_{0{,}975}\ \mathrm{SE}(\hat{\beta}_j)),\ \exp(\hat{\beta}_j + z_{0{,}975}\ \mathrm{SE}(\hat{\beta}_j))\Big).$$

2.5 Modelo Aditivo Generalizado — Regressão Logística (GAM-LR)

Os modelos aditivos generalizados (*generalized additive models* - GAM), propostos por Hastie e Tibshirani (1986, 1990), podem ser vistos como uma extensão dos modelos lineares generalizados, em que o preditor linear passa a incluir somas de funções suaves das covariáveis. Em termos gerais, um GAM assume a seguinte estrutura:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{q} \beta_j Z_{ij} + f_1(X_{1i}) + f_2(X_{2i}) + f_3(X_{3i}, X_{4i}) + \cdots,$$

onde $\mu_i = E(Y_i)$, $g(\cdot)$ é uma função de ligação suave e monótona, Z_{ij} representam covariáveis tratadas de forma paramétrica (tipicamente categóricas ou contínuas com efeito linear), β_j é o j-ésimo coeficiente linear, $f_k(\cdot)$ são funções suavizadas. De forma similar ao GLMs, supõe-se que as respostas Y_i sejam independentes e que

$$Y_i \sim \text{extensão da família exponencial}, \quad i = 1, \ldots, n.$$

O GAM com função de ligação logito (GAM-LR) estendem os modelos lineares generalizados ao permitir que parte dos efeitos das covariáveis seja modelada por funções suaves, de forma não paramétrica. Assim, enquanto alguns efeitos permanecem lineares,

outros podem apresentar comportamento não linear. De modo geral, o modelo logístico aditivo pode ser escrito como (HASTIE, 1986; WOOD, 2017):

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{q} \beta_j Z_{ij} + \sum_{k=1}^{r} f_k(X_{ik}),$$

onde Z_{ij} representam covariáveis tratadas de forma paramétrica (tipicamente categóricas ou contínuas com efeito linear) e $f_k(\cdot)$ são funções suavizadas.

Para estimar cada função suave $f_k(x)$, utiliza-se uma representação por funções de base. Assim, o suavizador é aproximado por uma combinação linear dessas bases:

$$f_k(x) = \sum_{m=1}^{M_k} b_{km}(x)\, \theta_{km} = \mathbf{b}_k(x)^\top \theta_k,$$

onde: $k = 1, \ldots, r$ identifica cada termo suave do modelo; M_k é o número de funções de base utilizadas para o termo k; $b_{km}(x)$ é a m-ésima função de base (como splines cúbicos, P-splines, B-splines ou thin-plate splines); θ_{km} é o coeficiente associado a $b_{km}(x)$; $fb_k(x)$ é o vetor das funções de base avaliadas em x; θ_k é o vetor de coeficientes do suavizador f_k.

Como funções muito flexíveis podem superajustar os dados, impõe-se uma penalização que controla a suavidade dos termos. A forma geral da penalização é dada por:

$$\mathcal{P} = \sum_{k=1}^{r} \lambda_k\, \theta_k^\top S_k\, \theta_k,$$

onde λ_k é o parâmetro de suavização do termo k; S_k é a matriz de penalização (rugosidade) associada ao termo k;$\theta_k^\top S_k \theta_k$ mede a complexidade da curva, penalizando oscilações excessivas.

A estimação simultânea dos coeficientes paramétricos e dos coeficientes das funções suaves é obtida a partir da maximização da *log-verossimilhança penalizada*. Para um modelo GAM-LR, seja

$$\ell(\beta, \theta)$$

a log-verossimilhança do modelo logístico padrão, dependente dos coeficientes paramétricos β e dos coeficientes das bases dos suavizadores θ. A log-verossimilhança penalizada é então dada por:

$$\ell_p(\beta, \theta) = \ell(\beta, \theta) - \mathcal{P},$$

onde $\mathcal{P}$ representa o termo de penalização aplicado às funções suaves.

A penalização introduz um termo que controla a complexidade dos suavizadores. A forma geral da log-verossimilhança penalizada é dada por:

$$\hat{\beta}, \hat{\theta} = \arg\max_{\beta,\theta} \left\{ \ell(\beta, \theta) - \frac{1}{2} \sum_{k=1}^{r} \lambda_k \theta_k^\top S_k \theta_k \right\}.$$

onde $\ell(\beta, \theta)$ é a log-verossimilhança do modelo logístico, sem penalização; r é o número de termos suaves no modelo; θ_k é o vetor de coeficientes das funções de base do termo suave $f_k(x)$; S_k é a matriz de penalização (ou matriz de rugosidade) associada ao termo k, construída a partir das derivadas de ordem superior do suavizador (por exemplo, integrando o quadrado da segunda derivada); λ_k é o parâmetro de suavização do termo k, controlando o balanço entre ajuste ao dado e suavidade: valores altos de λ_k impõem curvas mais suaves.

Os coeficientes paramétricos β_j são interpretados de maneira similar ao GLM-LR tradicional, em termos de mudanças no logito ou na razão de chances associada à mudança de x_a para x_b pode ser expressa por

$$\exp(f_k(x_b) - f_k(x_a)).$$

2.6 Seleção de Variáveis nos Modelos

A seleção de variáveis é uma etapa fundamental para garantir modelos com bom desempenho preditivo. Neste trabalho, foram utilizados procedimentos distintos para o GLM-LR e para o GAM-LR, respeitando as características de cada abordagem de modelagem.

2.6.1 Seleção de Variáveis no GLM-LR

No modelo GLM-LR foi aplicado o procedimento stepwise por meio da função stepAIC(), disponível no pacote MASS. Esse método busca o modelo com melhor equilíbrio entre ajuste e complexidade, utilizando como critério principal o AIC (Akaike Information Criterion). O AIC é definido por:

$$AIC = -2\log(L) + 2k,$$

em que L representa a verossimilhança do modelo e k é o número de parâmetros estimados. Modelos com menores valores de AIC são preferidos, pois indicam melhor qualidade de ajuste sem adicionar complexidade desnecessária.

Como medida complementar, também foi considerado o BIC (Bayesian Information Criterion), calculado por:

$$BIC = -2\log(L) + k\log(n),$$

que contém penalização mais rígida sobre o número de parâmetros. Assim, o procedimento adotado foi a seleção stepwise na direção both (adição e remoção); critério principal de escolha baseado no AIC e avaliação adicional do BIC para verificar a parcimônia das variáveis selecionadas. Esse procedimento permitiu identificar o conjunto mais adequado de variáveis para a composição final do GLM-LR.

2.6.2 Seleção de Variáveis no GAM-LR

Para o modelo GAM-LR, foi utilizado o método de penalização extra às funções suavizadoras, permitindo que variáveis irrelevantes tenham seus efeitos reduzidos a zero. Dessa forma, o próprio processo de estimação decide quais termos suaves devem permanecer no modelo final.

As principais vantagens dessa abordagem é evitar sobreajuste, selecionar automaticamente apenas efeitos significativos e manter a flexibilidade característica do GAM sem inflar o número de parâmetros. Assim, o GAM-LR resultará em um modelo enxuto, com efeitos suavizados apenas para as variáveis que apresentam contribuição relevante para a variabilidade do desfecho.

Este metodo está implementado no pacote *mgcv*, onde é possível ajustar automaticamente o modelo e realizar a seleção das variáveis relevantes utilizando o argumento $select = TRUE$ na função *gam*.

2.7 Predição e Performance

Uma vez ajustado o modelo de regressão logística, a probabilidade predita para cada observação i é dada por:

$$\hat{\pi}_i = \Pr(Y_i = 1 \mid X_i) = \frac{e^{X_i^\top \hat{\beta}}}{1 + e^{X_i^\top \hat{\beta}}}.$$

A partir dessa probabilidade, realiza-se a classificação binária por meio de um limiar (*threshold*) $t \in [0, 1]$:

$$\hat{Y}_i(t) = \begin{cases} 1, & \text{se } \hat{\pi}_i > t, \\ 0, & \text{caso contrário.} \end{cases}$$

A escolha do limiar afeta diretamente o desempenho preditivo do modelo. Para avaliar essa performance, utilizam-se medidas baseadas na matriz de confusão, cujas quantidades fundamentais são (veja tabela 1): (VP) verdadeiros positivos; (VN): verdadeiros negativos;(FP): falsos positivos;(FN) falsos negativos e Correlação de Matthews. Também foi utilizada a área sob a curva ROC.

Tabela 1 – Matriz de confusão

		Valor predito	
		1	**0**
Valor observado	**1**	VP (Verdadeiro positivo)	FN (Falso negativo)
	0	FP (Falso positivo)	VN (Verdadeiro negativo)

Fonte: Elaborado pela autora.

Assim descreveremos a seguir todas as métricas de desempenho empregadas posteriormente neste trabalho:

- Acurácia é a proporção de classificações corretas:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}.$$

- Sensibilidade mede a capacidade do modelo em identificar corretamente o evento 1 (positivo):

$$SE = \frac{VP}{VP + FN}.$$

- Especificidade quantifica a capacidade do modelo em identificar corretamente o evento 0 (negativo):

$$SP = \frac{VN}{VN + FP}.$$

- Valor Preditivo Positivo (VPP) representa a proporção de predições positivas que são corretas:

$$VPP = \frac{VP}{VP + FP}.$$

- Valor Preditivo Negativo (VPN) indica a proporção de predições negativas corretas:

$$VPN = \frac{VN}{VN + FN}.$$

- Coeficiente de Matthews (MCC) é uma medida global de correlação entre predições e valores verdadeiros, variando de -1 (desempenho oposto ao perfeito) a $+1$ (classificação perfeita):

$$MCC = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}.$$

- Área sob a Curva ROC (AUC) quantifica a discriminação do modelo, isto é, sua capacidade de atribuir probabilidades maiores aos casos positivos do que aos negativos. A curva ROC é construída variando-se o limiar t, e o AUC corresponde à integral sob essa curva.

2.8 Validação Cruzada

A validação cruzada do tipo leave-one-out(LOOCV) é uma técnica amplamente utilizada para avaliar o desempenho preditivo de modelos estatísticos. Nesse procedimento, cada observação é retirada do conjunto de dados uma vez e utilizada como conjunto de teste individual. Para um conjunto contendo n observações, realizam-se, portanto, n predições de teste do modelo.

Para cada modelo, o procedimento consiste nos seguintes passos:

- Remoção da observação i: Ajusta-se o modelo utilizando todas as observações exceto a i.
- Predição no conjunto de teste: Com o modelo ajustado, calculam-se a probabilidade predita $\hat{\pi}_i$.
- Seleção do cutoff ideal no treino: Uma vez encontradas todas as probabilidades $\hat{\pi}_i$ para $i = 1, ..., n$ o conjunto de treino, o limiar de classificação é escolhido buscando o valor que maximiza o coeficiente de Matthews (MCC):

$$t = \arg\max_{0<k<1} \text{MCC}(k).$$

- Classificação: O cutoff t é utilizado juntamente com as probabilidades $\hat{\pi}_i$ para definir a classe predita:

$$\hat{Y}_i^{LOO} = 1\{\hat{\pi}_i > t\}.$$

Nesse procedimento:$\hat{\pi}_i$ é a probabilidade predita da observação i usando o modelo ajustado sem ela; t é o cutoff ótimo e$\hat{Y}_i^{LOO}$ é a classe atribuída à observação i pelo processo de LOOCV.

A aplicação do LOOCV reduz o viés da avaliação preditiva, pois cada observação é prevista por um modelo ajustado sem seu próprio valor, evitando sobreajuste. Ao final, as classificações $\hat{Y}_i^{LOO}$ são comparadas com os valores verdadeiros para calcular todas as métricas de desempenho utilizadas neste trabalho (acurácia, sensibilidade, especificidade, MCC e AUC).

3 Resultados

3.1 Estatística Descritiva

Conforme mencionado na introdução, utilizamos os dados do projeto SanRural realizado no estado de Goiás no período 2018 á 2019. Após a etapa de limpeza, foram analisadas 502 observações. O desfecho analisado foi a presença de triatomíneos, registrada pela variável pos01 na tabela de dados, definida como 1 = presença do vetor e 0 = ausência. A prevalência de triatomíneos foi de 10.7%, indicando que a maioria das comunidades amostradas não apresentou ocorrência do vetor.

3.2 Estatística Descritiva das Variáveis Categóricas

A análise descritiva das variáveis categóricas permite compreender o contexto social, territorial e ambiental das famílias incluídas no estudo. As comunidades avaliadas apresentam distribuição relativamente equilibrada, com cada localidade contribuindo entre 1% e 6% da amostra total.

Essa diversidade também se expressa nos municípios envolvidos. Embora a amostra esteja distribuída por várias regiões, destaca-se o município de Silvânia, que reúne quase 10% dos participantes. Os municípios Barro Alto e Santa Rita do Novo Destino também se sobressaem, com participações de 8.12% e 4.95%, respectivamente. Os demais municípios apresentam frequências menores, mas compõem um cenário territorial relevante para o estudo.

A Tabela 2 apresenta a distribuição das principais variáveis categóricas avaliadas no estudo.

Tabela 2 – Estatística descritiva das variáveis categóricas

Variável	**Categoria**	**Freq. (n)**	**(%)**
Tipologia da comunidade	Assentamento	186	36.83
	Quilombola	301	59.60
	Ribeirinho	18	3.56
Parede médio/alto risco	Não	369	73.07
	Sim	136	26.93
Água em caixa de amianto	Não	444	87.92
	Sim	61	12.08
Presença de triatomíneos	Não	451	89.31
	Sim	54	10.69
Mosquiteiro	Não	292	57.82
	Sim	213	42.18
Qtd. positivo DC prévio	Não	468	92.67
	Sim	37	7.33

Fonte: Elaborado pela autora.

A tipologia das comunidades revela um aspecto marcante da realidade investigada: a maior parte da amostra é formada por territórios Quilombolas 59.60%, seguidos por Assentamentos Rurais 36.83% e, em menor proporção, comunidades Ribeirinhas 3.56%. Essa composição evidencia a forte presença de populações tradicionais, frequentemente expostas a condições de vulnerabilidade social e ambiental que podem influenciar diretamente os fatores de risco analisados.

No âmbito das condições domiciliares, verificou-se que 26.93% das residências apresentam paredes classificadas como de médio ou alto risco, o que pode favorecer a presença de vetores ou comprometer a integridade estrutural do domicílio. Além disso, 12.08% das famílias relataram utilizar caixa d'água de amianto para atividades domésticas, prática que possui reconhecidos riscos à saúde.

A presença de triatomíneos, vetores da Doença de Chagas, foi relatada por 10.69% dos domicílios. Embora esse percentual não represente a maioria, ele é epidemiologicamente relevante e reforça a importância da vigilância entomológica nas áreas estudadas.

Em relação às medidas de proteção individual, observou-se que 42.18% das famílias utilizam mosquiteiros, um recurso simples e eficaz para reduzir o contato com vetores. Entretanto, a maioria 57.82% não faz uso desse equipamento, o que pode aumentar a vulnerabilidade às picadas de insetos, especialmente em regiões endêmicas.

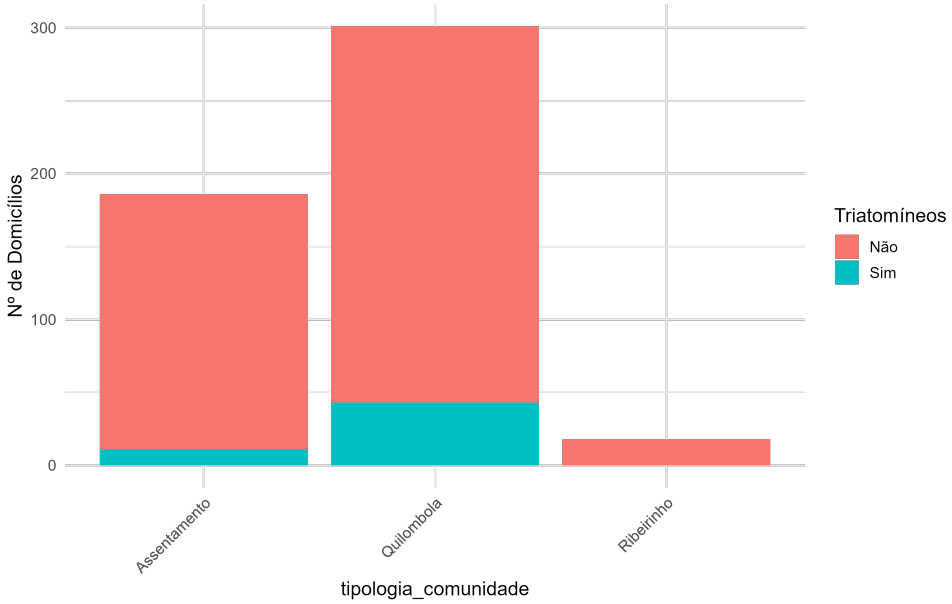
Quanto à positividade prévia para Doença de Chagas (DC) no domicílio, 7.33% relataram pelo menos um caso positivo. Esse indicador sugere exposição histórica ao vetor e destaca a necessidade de ações direcionadas de vigilância e prevenção nas comunidades.

De maneira geral, os resultados apontam um cenário marcado pela diversidade territorial, forte presença de povos e comunidades tradicionais e variações significativas

nas condições domiciliares e ambientais. Esses elementos oferecem subsídios essenciais para a interpretação dos modelos estatísticos apresentados nas seções posteriores e contribuem para uma compreensão mais ampla dos fatores associados ao risco de infecção na população estudada.

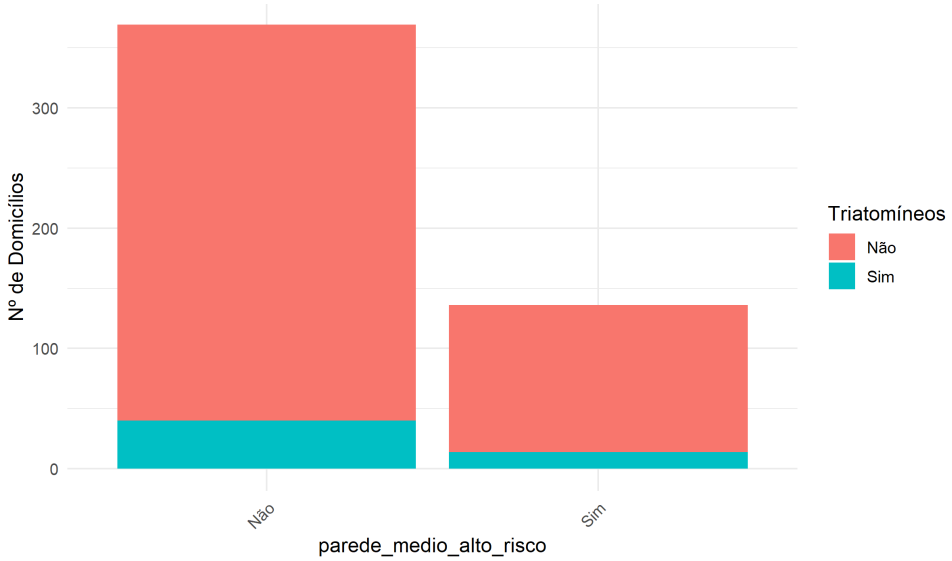
A seguir, apresentam-se de forma ilustrativa os gráficos correspondentes às principais variáveis categóricas analisadas, com o objetivo de complementar a leitura dos resultados descritivos.

Figura 1 – Gráfico da Tipologia da Comunidade



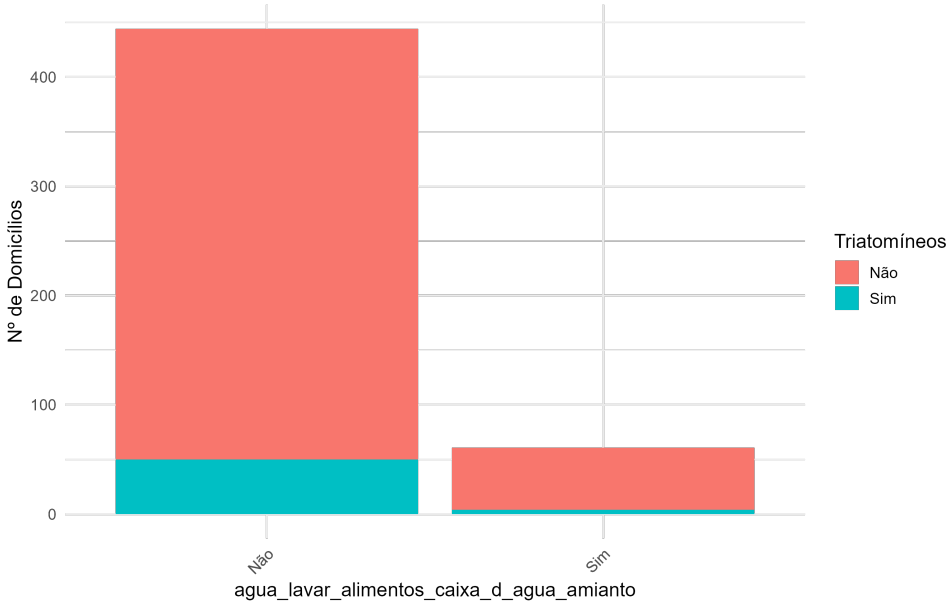
Fonte: Elaborado pela autora.

Figura 2 – Gráfico do tipo de parede



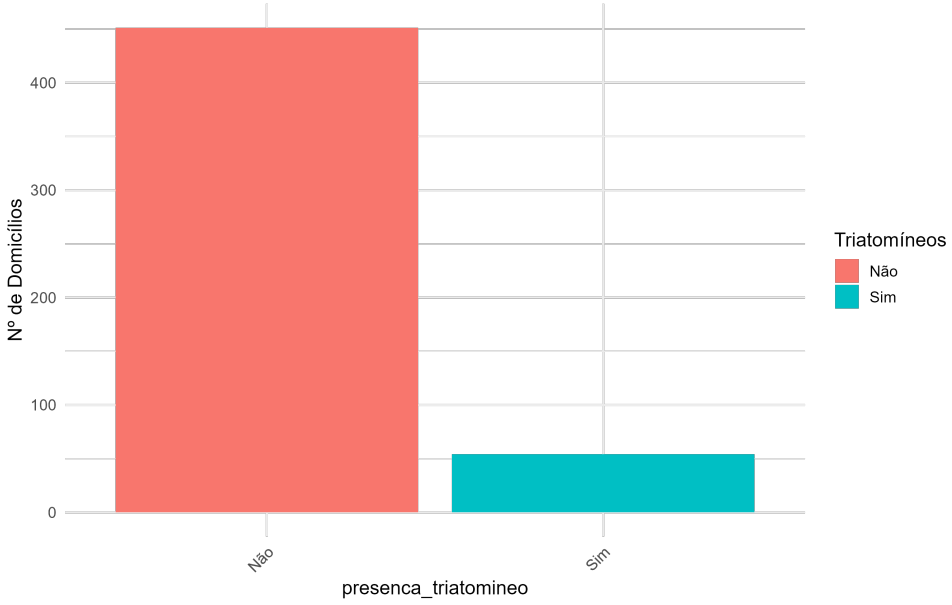
Fonte: Elaborado pela autora.

Figura 3 – Gráfico do tipo de caixa de água



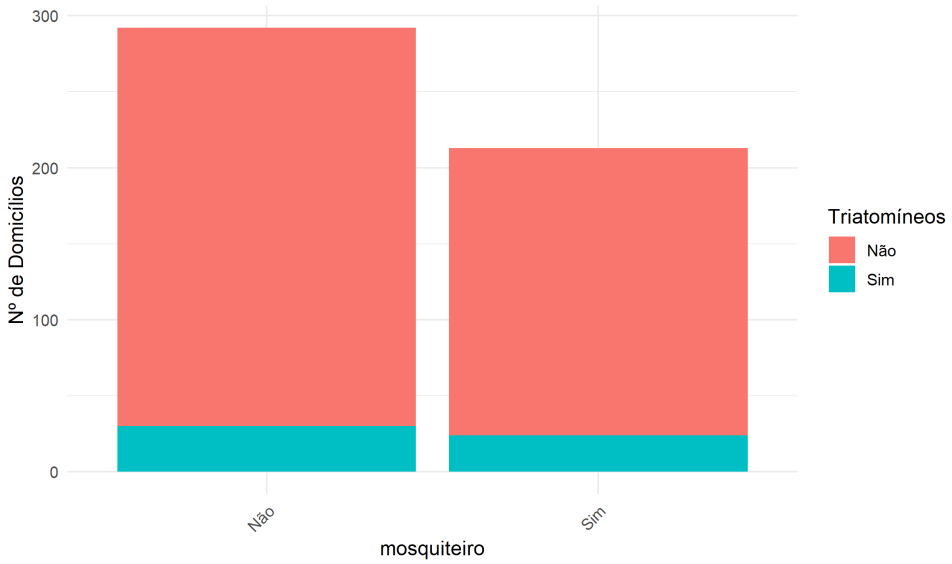
Fonte: Elaborado pela autora.

Figura 4 – Gráfico da presença de triatomíneos



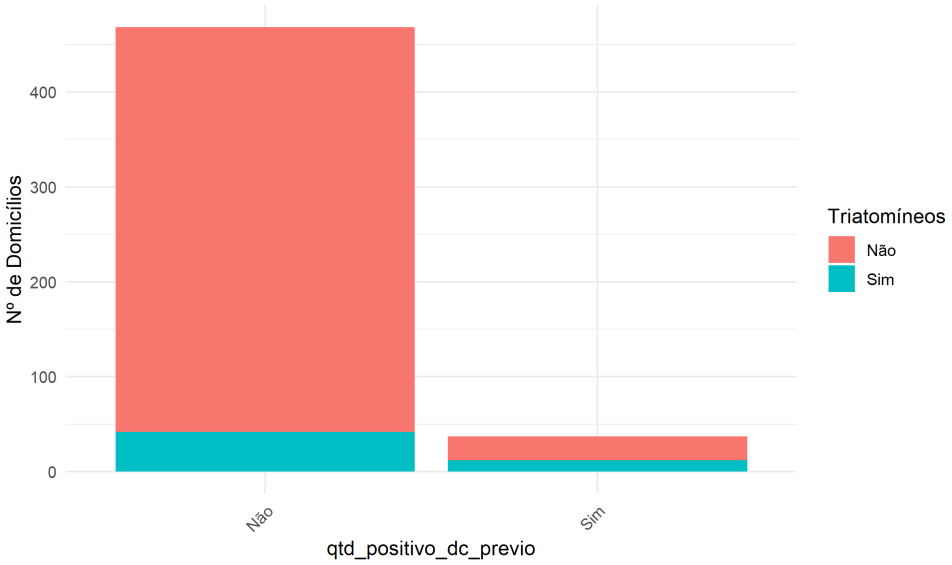
Fonte: Elaborado pela autora.

Figura 5 – Gráfico do uso mosquiteiro



Fonte: Elaborado pela autora.

Figura 6 – Gráfico da Qtd. positivo DC prévio



Fonte: Elaborado pela autora.

3.3 Estatística Descritiva das Variáveis Contínuas

A Tabela 3 apresenta as estatísticas descritivas das variáveis contínuas analisadas no estudo, oferecendo uma visão geral das características das residências e da presença de animais nos domicílios investigados.

Tabela 3 – Estatística descritiva das variáveis contínuas

Variável	**Média**	**Mediana**	**DP**	**Mínimo**	**Máximo**
Quantidade de porcos confinados	3.71	1	6.40	0	60
Quantidade de porcos soltos	0.26	0	1.49	0	13
Quantidade de currais	11.93	0	25.19	0	200
Galinhas no galinheiro	6.59	0	19.51	0	250
Galinhas soltas	27.33	20	32.96	0	200
Animais de estimação	3.19	3	3.04	0	18
Moradores 0 a 1 ano	0.07	0	0.27	0	2
Moradores 2 a 12 anos	0.43	0	0.83	0	6
Moradores 13 a 18 anos	0.31	0	0.63	0	4
Moradores 19 a 59 anos	1.61	2	1.09	0	6
Moradores 60 anos ou mais	0.63	0	0.78	0	3
Total de residentes	3.06	3	1.56	1	10
Homens na residência	1.65	1	1,02	0	6
Mulheres na residência	1.41	1	0.97	0	6
Total de cômodos	6.34	6	1.67	1	14
Relação residentes/quartos	1.16	1	0.73	0.2	7
Idade do responsável pelo domicílio	51.6	53	15.71	18	91

Fonte: Elaborado pela autora

A análise das estatísticas revela uma considerável heterogeneidade na presença de animais nos domicílios, refletida pelos elevados desvios-padrão em várias variáveis. No caso dos porcos confinados, observa-se uma média de 3.71 animais por residência, mas a mediana é apenas 1, indicando que a maior parte dos domicílios possui poucos animais, enquanto alguns apresentam concentrações elevadas, chegando a 60 exemplares. Situação similar é observada para as galinhas mantidas em galinheiro (média 6.59; mediana 0; máximo 250), evidenciando que apenas alguns domicílios concentram grandes quantidades desses animais. Já as galinhas soltas demonstraram maior frequência e distribuição mais uniforme entre os domicílios (média 27.33; mediana 20).A idade do responsável pelo domicílio apresentou média de 51.6 anos, mediana de 53 anos, com desvio-padrão de 15.71 anos, variando entre 18 e 91 anos.

Os animais de estimação apresentam uma distribuição mais equilibrada, com média de 3.19 e mediana de 3, sugerindo homogeneidade relativa entre os domicílios. Quanto às variáveis demográficas, nota-se baixo número de crianças pequenas, adolescentes indicando menor dispersão.

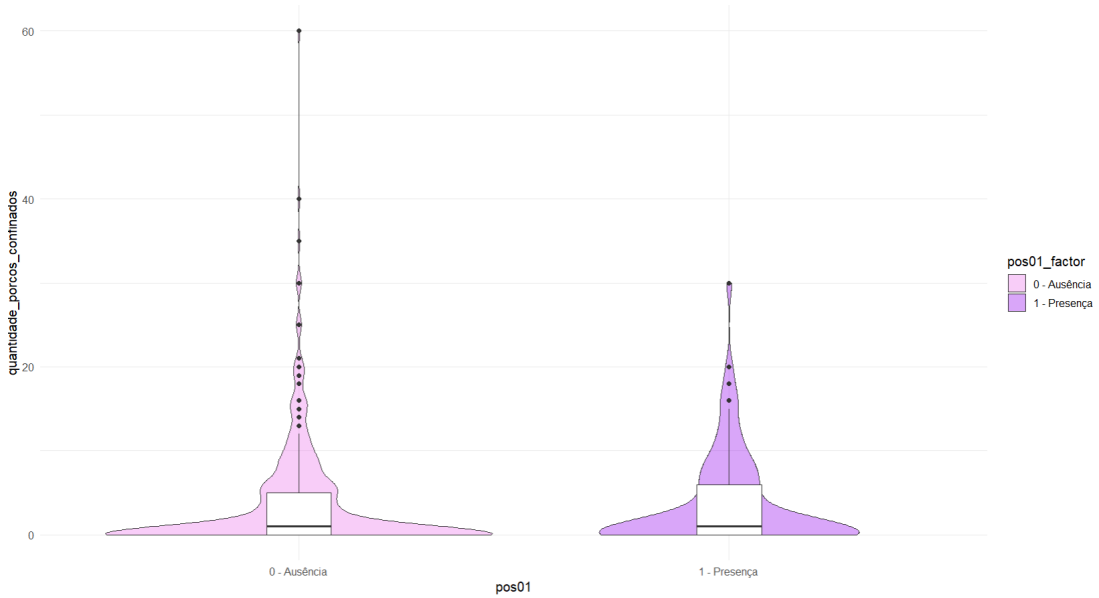
De maneira geral, os resultados evidenciam diferenças marcantes tanto na criação de animais quanto na composição familiar, aspectos que podem influenciar diretamente a dinâmica de risco de infestação por triatomíneos nas comunidades estudadas. Essas variações reforçam a importância de considerar o contexto domiciliar e a distribuição de animais ao planejar estratégias de vigilância e controle da doença.

O gráfico complementar à Tabela 3 facilita a visualização da distribuição das variáveis contínuas, permitindo identificar padrões que os números resumidos não captam completamente. Observa-se que variáveis como o número de porcos confinados e de galinhas em galinheiro apresentam distribuições altamente assimétricas, com a maioria dos domicílios concentrando-se em valores próximos a zero, mas com alguns casos extremos outliers de grande magnitude.

No caso das galinhas soltas, a dispersão é mais equilibrada, indicando maior frequência e homogeneidade dessa prática entre as famílias avaliadas. A presença de animais de estimação apresenta distribuição relativamente homogênea, concentrando a maioria dos domicílios em até cinco animais. Já as variáveis demográficas, como número de crianças, adolescentes evidenciam baixa variabilidade, sugerindo uniformidade nesses aspectos entre os domicílios.

Logo a seguir vamos ver os gráficos tipo violino combinados com boxplot descrevendo visualmente as estatisticas das variaveis estudadas no qual a criação de animais, especialmente galinhas soltas, é uma prática central e potencialmente relevante para a exposição estudada, enquanto outras variáveis contribuem menos para a heterogeneidade observada nos domicílios.

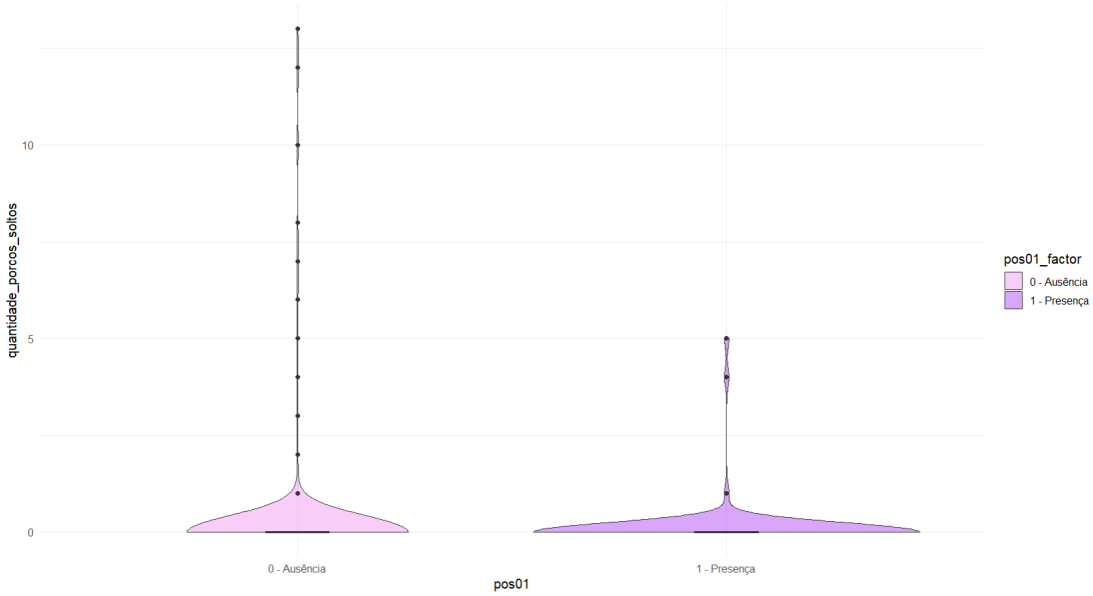
Figura 7 – Boxplot quantidade de porcos confinados



Fonte: Elaborado pela autora

O boxplot mostra que a maioria dos domicílios possui pequena quantidade de porcos confinados, com alguns domicílios apresentando valores elevados, caracterizando outliers.

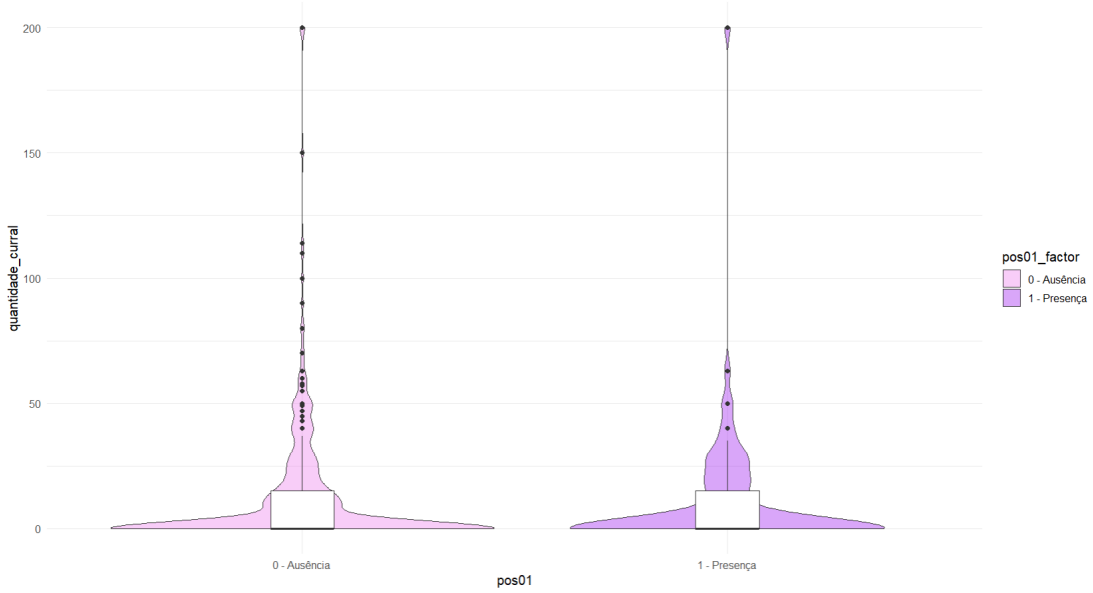
Figura 8 – Boxplot quantidade de porcos soltos



Fonte: Elaborado pela autora

A distribuição do número de porcos soltos é assimétrica, com poucos domicílios mantendo altas quantidades, sugerindo prática heterogênea de criação extensiva.

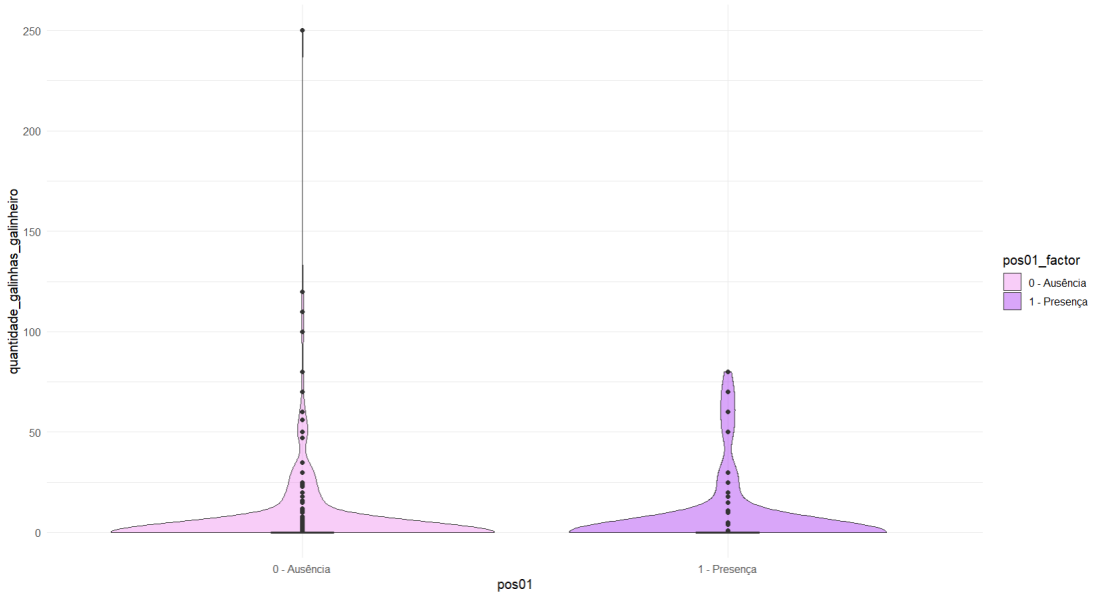
Figura 9 – Boxplot quantidade de curral



Fonte: Elaborado pela autora

A maioria dos domicílios apresenta poucos currais, com alguns domicílios possuindo quantidade significativamente maior, refletindo diversidade estrutural na criação animal.

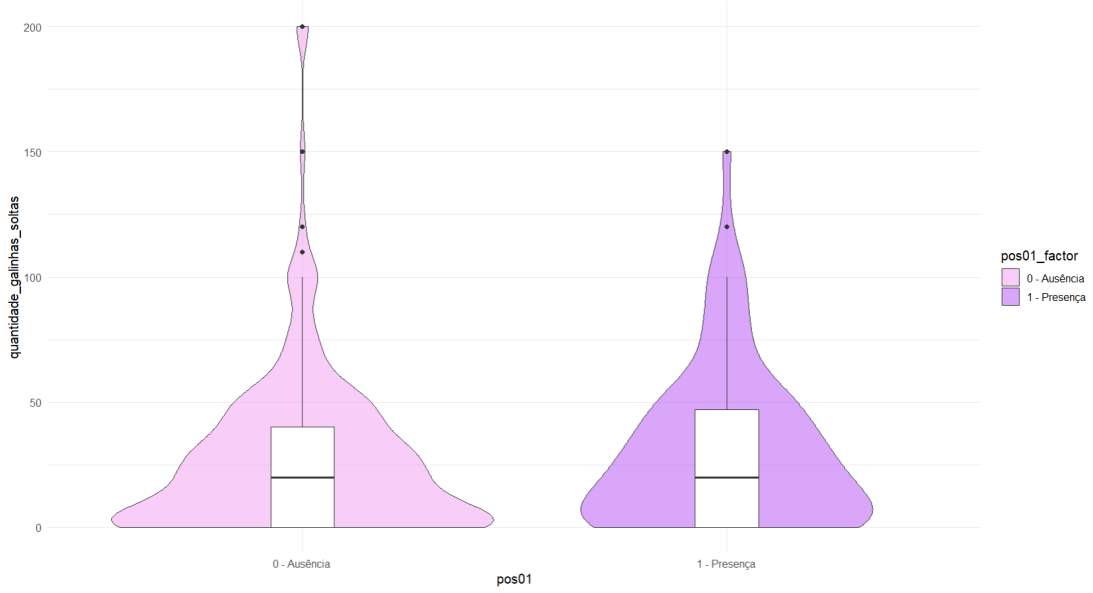
Figura 10 – Boxplot quantidade de galinhas no galinheiro



Fonte: Elaborado pela autora

O boxplot indica que grande parte dos domicílios mantém pequenas quantidades de galinhas em galinheiro, com alguns domicílios apresentando valores mais elevados, configurando outliers.

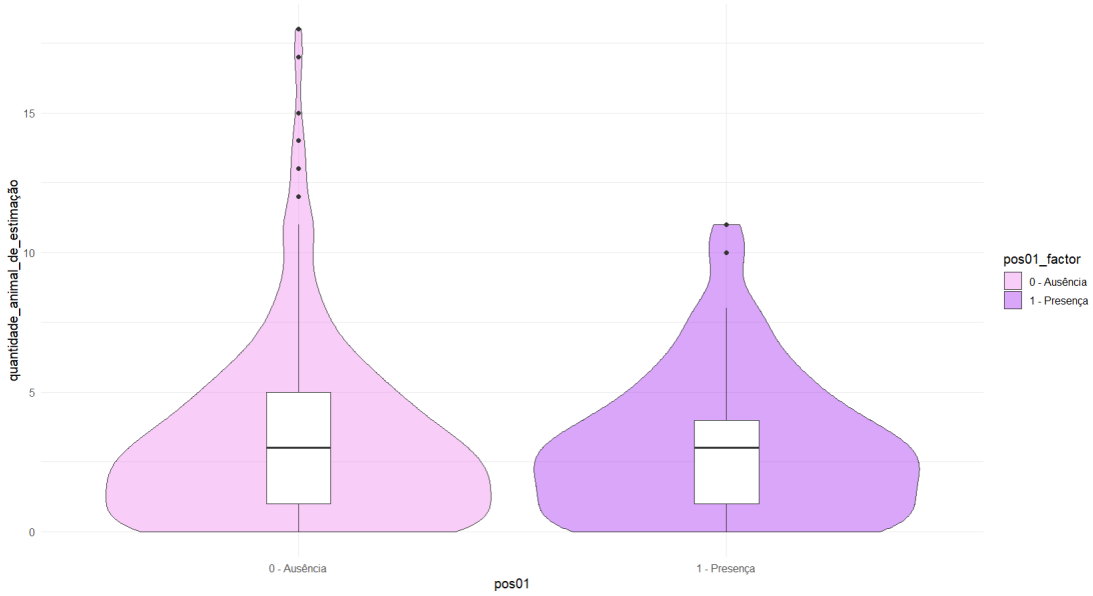
Figura 11 – Boxplot quantidade de galinhas soltas



Fonte: Elaborado pela autora

A distribuição do número de galinhas soltas é mais dispersa, com alguns domicílios apresentando grandes quantidades, evidenciando heterogeneidade na criação extensiva.

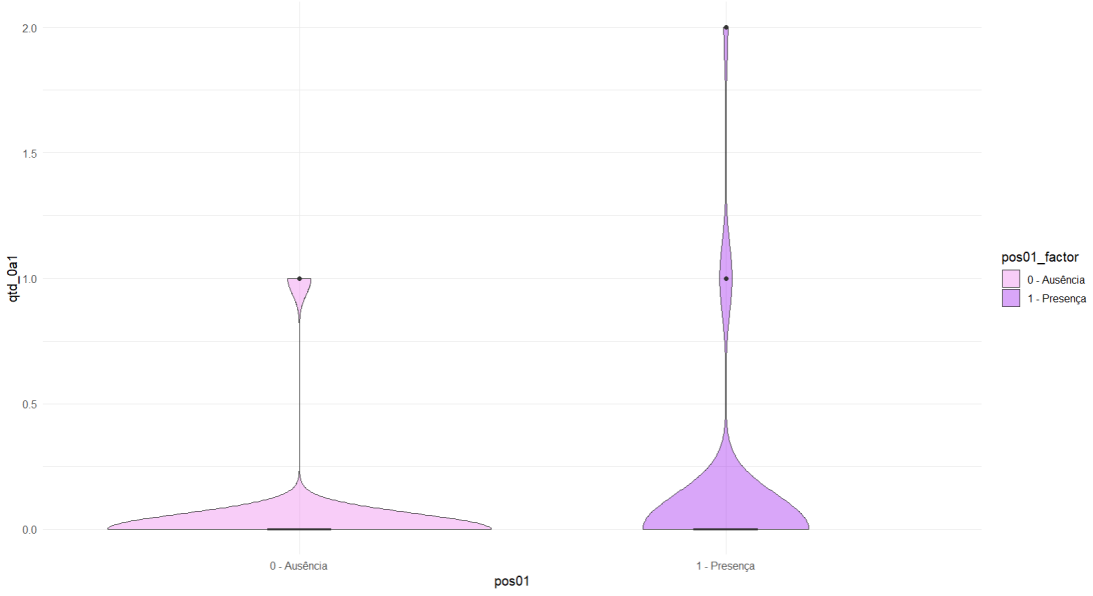
Figura 12 – Boxplot quantidade de animais de estimação



Fonte: Elaborado pela autora

A maioria dos domicílios apresenta poucos animais de estimação, com poucos outliers indicando casas com maior número de animais domésticos.

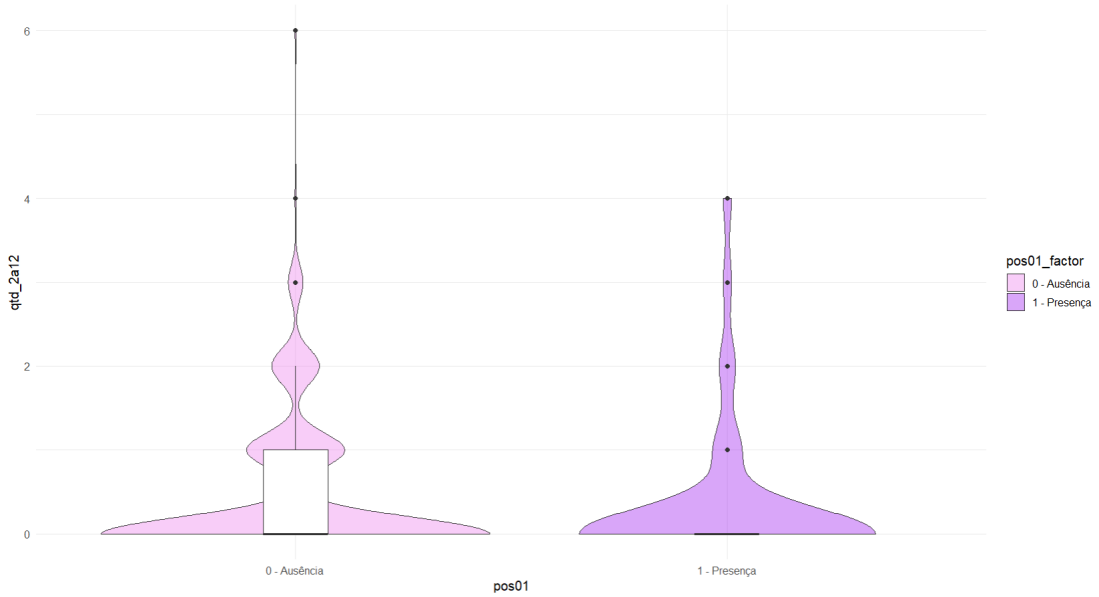
Figura 13 – Boxplot quantidade de crianças de 0 a 1 anos



Fonte: Elaborado pela autora

A distribuição indica que a maioria dos domicílios possui poucas crianças nesta faixa etária, com poucos outliers representando famílias maiores com bebês.

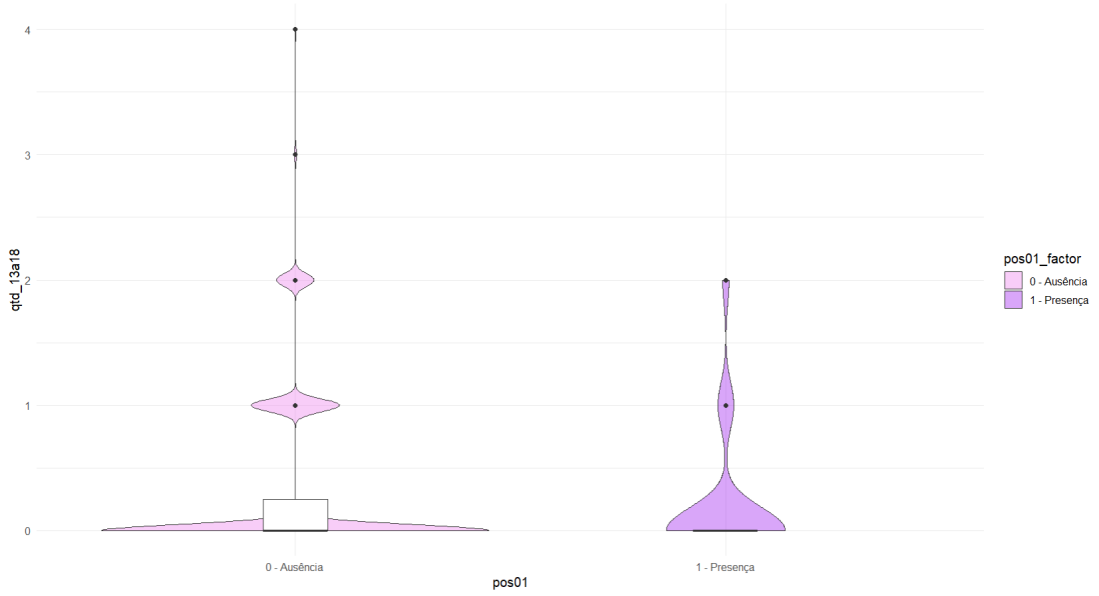
Figura 14 – Boxplot quantidade de crianças de 2 a 12 anos



Fonte: Elaborado pela autora

O número de crianças de 2 a 12 anos é moderadamente disperso, com a maioria dos domicílios apresentando valores próximos à mediana e poucos outliers.

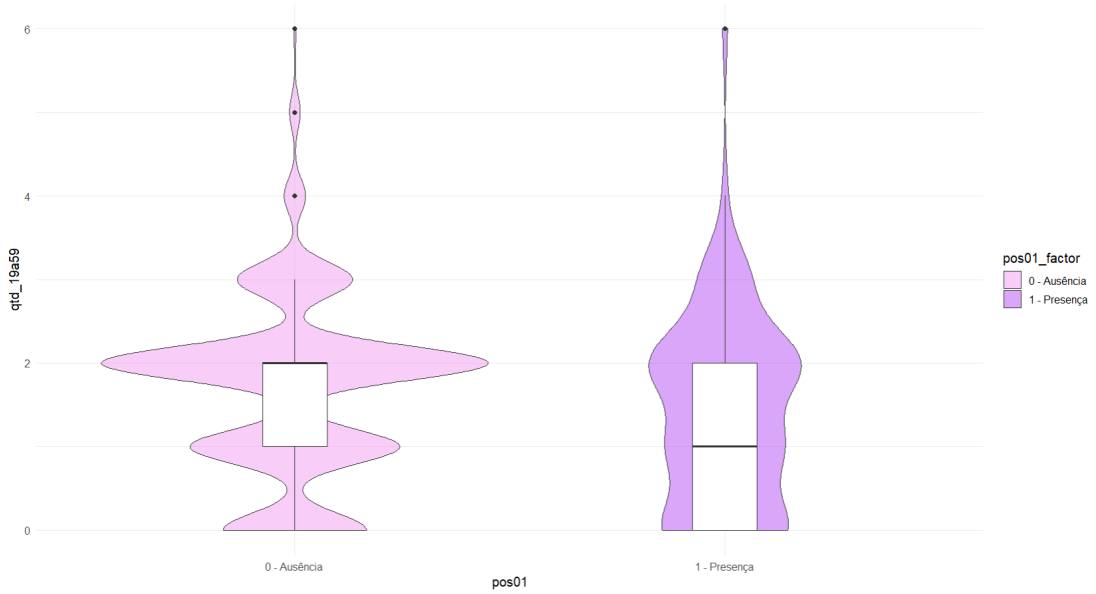
Figura 15 – Boxplot quantidade de adolescentes de 13 a 18 anos



Fonte: Elaborado pela autora

A distribuição mostra que poucos domicílios possuem adolescentes nesta faixa, e a maioria concentra-se próximo à mediana baixa, com presença de outliers.

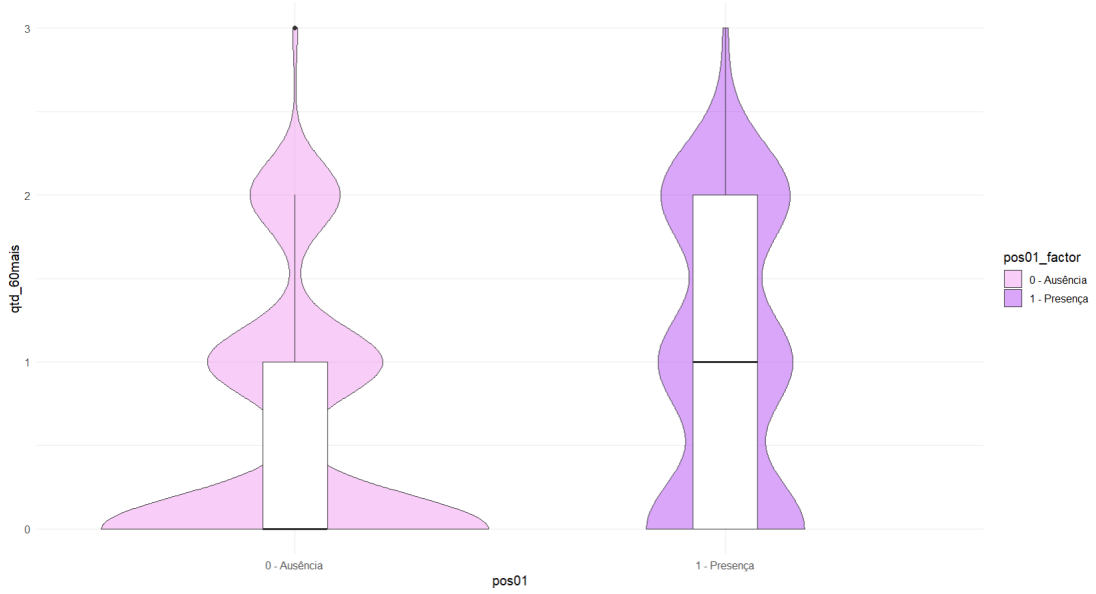
Figura 16 – Boxplot quantidade de adultos de 19 a 59 anos



Fonte: Elaborado pela autora

O boxplot evidencia que a maior parte dos domicílios possui adultos de 19 a 59 anos em quantidade média, com poucos domicílios concentrando número elevado.

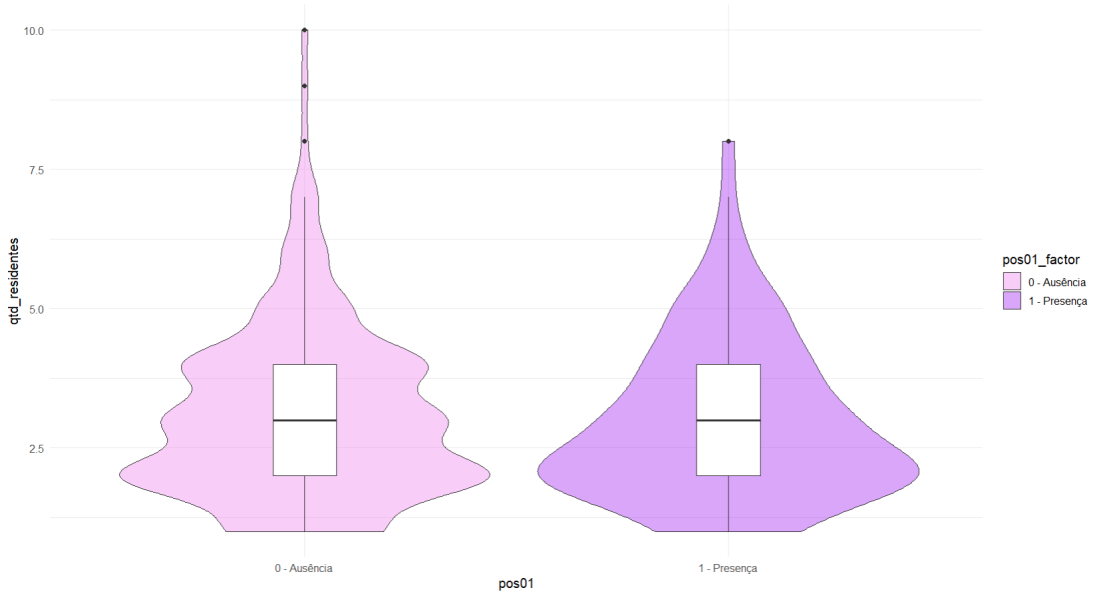
Figura 17 – Boxplot quantidade de idosos de 60 anos ou mais



Fonte: Elaborado pela autora

A maioria dos domicílios possui poucos idosos, refletindo baixa proporção desta faixa etária, com alguns domicílios contendo valores mais elevados.

Figura 18 – Boxplot relação de residentes por quartos



Fonte: Elaborado pela autora

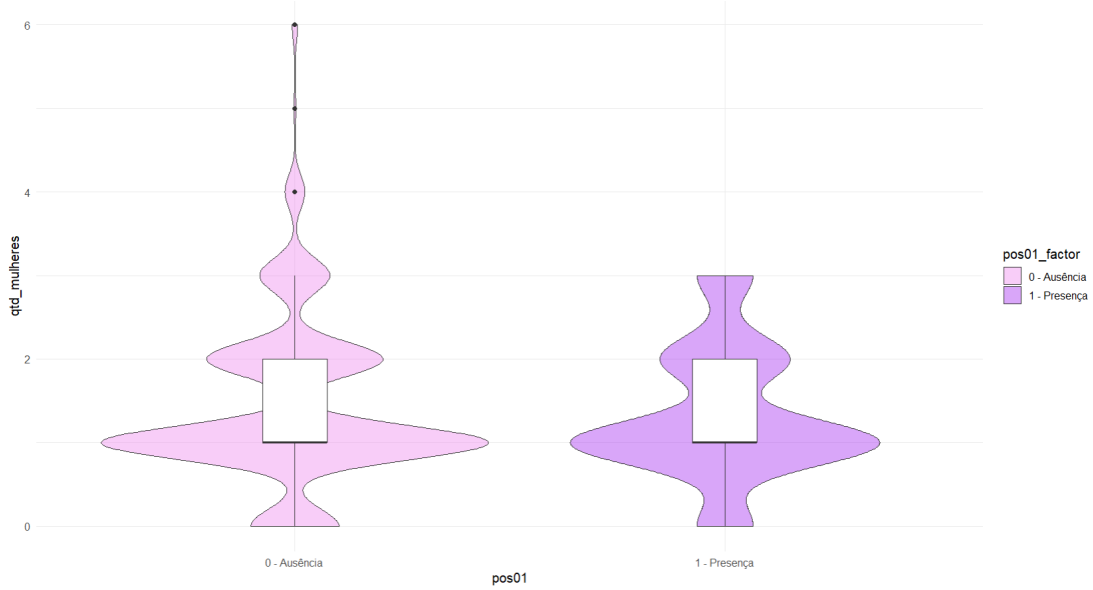
A relação residentes/quartos apresenta dispersão moderada, com maioria dos domicílios mantendo proporção equilibrada e alguns casos de superlotação.

Figura 19 – Boxplot quantidade de homens

Fonte: Elaborado pela autora

O boxplot mostra que a maioria dos domicílios possui quantidade equilibrada de homens, com poucos valores extremos.

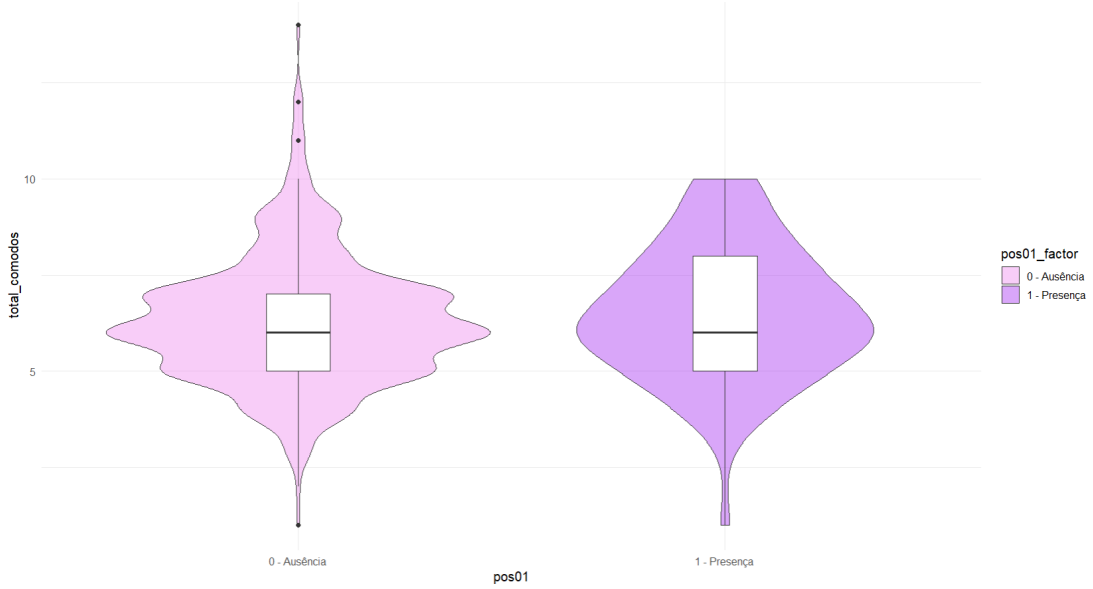
Figura 20 – Boxplot quantidade de mulheres



Fonte: Elaborado pela autora

A distribuição de mulheres é semelhante à de homens, concentrando-se em torno da mediana e apresentando poucos outliers.

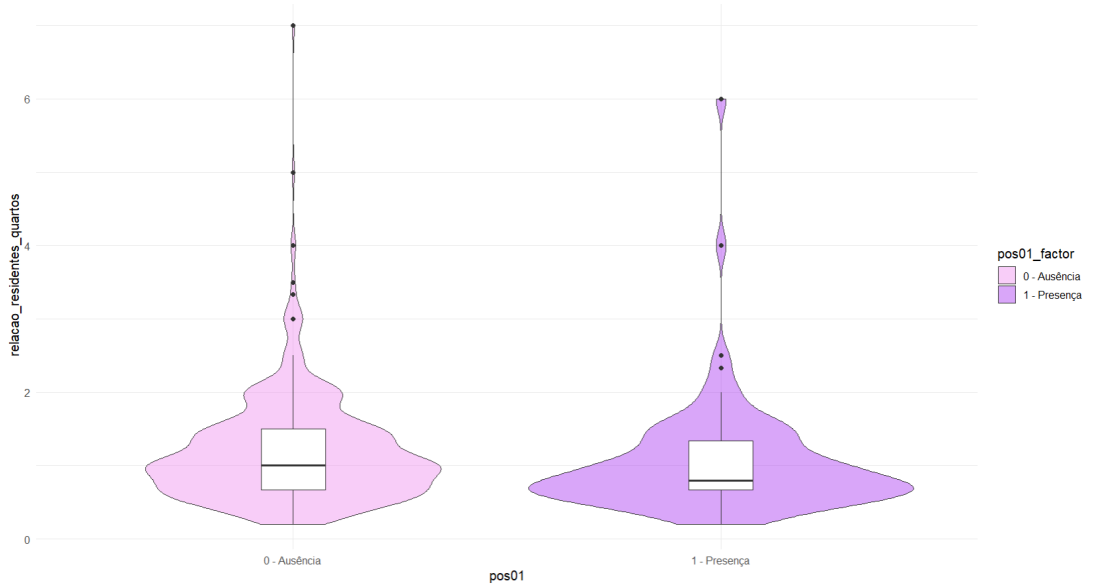
Figura 21 – Boxplot da distribuição do total de cômodos



Fonte: Elaborado pela autora

O formato do violino permite visualizar tanto a densidade dos valores quanto a dispersão, mostrando que a maior concentração de domicílios está em torno de 5 a 7 cômodos, com poucos domicílios extremos em ambas as pontas da distribuição.

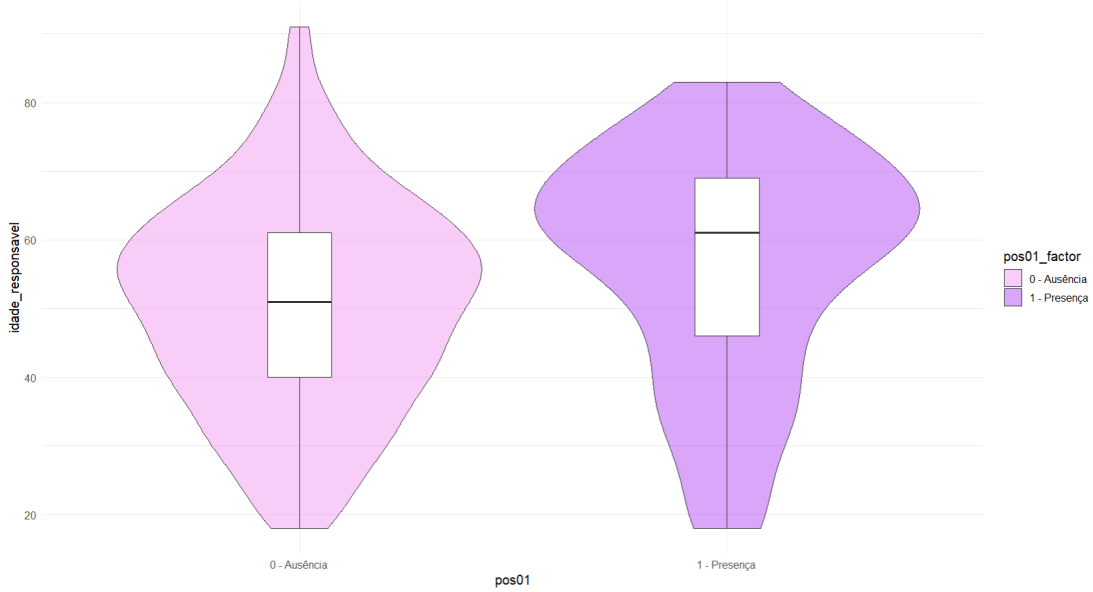
Figura 22 – Boxplot relação de residentes por quartos



Fonte: Elaborado pela autora

A relação residentes/quartos apresenta dispersão moderada, com maioria dos domicílios mantendo proporção equilibrada e alguns casos de superlotação.

Figura 23 – Boxplot relação da idade do responsável



Fonte: Elaborado pela autora

Os gráficos mostram maior concentração de indivíduos na faixa etária central, enquanto as extremidades mais estreitas representam as idades menos frequentes nas pontas da distribuição.

3.4 Resultados do Modelo Linear Generalizado - Regressão Logística (GLM-LR)

O modelo GLM foi inicialmente estimado com todas as covariáveis definidas no escopo da análise. Em seguida, aplicou-se o procedimento *stepwise* com critério AIC, permitindo seleção do melhor conjunto de variáveis por inclusão e remoção segundo o critério AIC. O modelo selecionado apresentou melhor desempenho segundo os critérios de informação, com redução do AIC e do BIC quando comparado ao modelo completo. O modelo selecionado reteve os seguintes preditores: tipologia, idade do responsável, número de crianças de 0 a 1 ano, positivo prévios para doença de Chagas, relação de moradores por quarto, parede de risco e água acumulada em caixa de amianto. Esses resultados indicam que, no ajuste linear generalizado, variáveis socioambientais e histórico epidemiológico permanecem como determinantes centrais da infecção onde apresentamos na Tabela 4.

Tabela 4 – Resultados do GLM-LR: coeficientes, erros-padrão e testes z.

Termo	**Estimativa**	**Erro-padrão**	**Valor Z**	**p-valor**
Intercepto	-3.840	0.712	-5.40	6.66×10^{-8}
Tipologia: Quilombola	1.170	0.325	3.60	3.17×10^{-4}
Tipologia: Ribeirinho	-0.475	1.130	-0.42	6.73×10^{-1}
Idade do responsável	0.0226	0.00926	2.44	1.47×10^{-2}
Qtd. crianças 0 a 1 ano	1.210	0.444	2.72	6.54×10^{-3}
Positivo prévio para DC	2.720	0.422	6.46	1.08×10^{-10}
Uso de mosquiteiro	-0.353	0.280	-1.26	2.07×10^{-1}
Relação moradores/quartos	-0.233	0.233	-1.00	3.16×10^{-1}
Parede médio/alto risco	0.609	0.286	2.13	3.35×10^{-2}
Uso de caixa d'água de amianto	0.636	0.372	1.71	8.76×10^{-2}

Fonte: Elaborado pela autora.

A Tabela 4 apresenta os coeficientes estimados do modelo logístico, incluindo erros-padrão, valores z e níveis de significância. O modelo foi estimado com 502 observações completas o AIC do modelo selecionado 394.78 e o BIC de 428.54.

A interpretação dos coeficientes segue a lógica do modelo logístico: valores positivos de $\hat{\beta}_j$ aumentam as chances do desfecho (OR > 1), enquanto valores negativos reduzem tais chances (OR < 1). A Tabela 5 apresenta as razões de chances estimadas e seus intervalos de confiança de 95% .

No modelo de regressão logística (GLM-LR), em que o desfecho foi definido como a presença de triatomíneos no domicílio, observou-se que a tipologia Quilombola apresentou maior chance de ocorrência do desfecho em comparação ao grupo de referência, com razão de chances (OR) de aproximadamente 3,22 (IC95%: 1,70–6,09). Isso significa que domicílios em comunidades Quilombolas apresentam, em média, pouco mais de três

Tabela 5 – Razões de chances (OR) e intervalos de confiança (IC 95%) do GLM-LR.

Variável	**OR**	**IC 2.5%**	**IC 97.5%**
Intercepto	0.0214	0.0053	0.0864
Tipologia da comunidade: Quilombola	3.2231	1.7047	6.0942
Tipologia da comunidade: Ribeirinho	0.6216	0.0682	5.6677
Idade do responsável	1.0228	1.0045	1.0416
Qtd. de moradores 0 a 1 ano	3.3444	1.4010	7.9841
Positividade prévia para DC	15.1981	6.6525	34.7215
Uso de mosquiteiro	0.7023	0.4058	1.2153
Relação residentes/quartos	0.7920	0.5020	1.2493
Parede de médio/alto risco	1.8377	1.0485	3.2209
Água de caixa de amianto	1.8896	0.9105	3.9213

Fonte: Elaborado pela autora

vezes a chance de presença de triatomíneos em relação às demais tipologias, evidenciando associação estatisticamente significativa.

A idade do responsável mostrou efeito positivo discreto, com OR de 1,02 (IC95%: 1,00–1,04), indicando que, a cada ano adicional de idade, a chance de presença de triatomíneos aumenta cerca de 2%, mantendo-se constantes as demais variáveis do modelo. De forma semelhante ao exemplo, o número de moradores de 0 a 1 ano também se associou a maior risco: para cada criança adicional nessa faixa etária, a razão de chances estimada foi de 3,34 (IC95%: 1,40–7,98), sugerindo aumento importante na probabilidade de infestação domiciliar.

A variável de maior magnitude foi a positividade prévia para doença de Chagas no domicílio, cuja OR foi de 15,20 (IC95%: 6,65–34,72). Esse resultado indica que domicílios com histórico prévio positivo apresentam cerca de quinze vezes a chance de presença atual de triatomíneos em comparação àqueles sem histórico, reforçando uma forte associação entre ocorrência anterior e infestação presente. Além disso, domicílios com paredes classificadas como de médio/alto risco apresentaram OR de 1,84 (IC95%: 1,05–3,22), o que corresponde a aproximadamente 84% de aumento na chance de presença de triatomíneos em relação às moradias com paredes de baixo risco, sugerindo que características construtivas desfavoráveis contribuem para maior vulnerabilidade à infestação.

Por outro lado, as variáveis tipologia Ribeirinho, uso de mosquiteiro, relação residentes/quartos e água armazenada em caixa d'água de amianto apresentaram intervalos de confiança que incluem o valor 1, indicando ausência de evidência estatística robusta de associação com o desfecho na amostra analisada. Embora algumas dessas estimativas apontem tendências plausíveis — como um possível efeito protetor do uso de mosquiteiro ($OR < 1$) e maior risco associado à água de caixa de amianto ($OR > 1$) — tais achados devem ser interpretados com cautela e idealmente confirmados em estudos futuros com

maior poder amostral.

3.5 Resultados do Modelo Aditivo Generalizado (GAM-LR)

No GAM-LR, a seleção de termos ocorre por penalização extra aplicada aos termos suavizadores, resduzindo complexidade do modelo. No pacote *mgcv* do software R, quando o parâmetro *select = TRUE* é ativado, suavizadores podem ter seus graus de liberdade efetivos reduzidos a valores próximos de zero, resultando na remoção prática do termo.

Na aplicação ao conjunto de dados, observou-se que o suavizador para histórico de positividade manteve-se altamente significativo; o efeito suave da *idade do responsável* permaneceu no modelo, mas com baixa complexidade;o suavizador para *número de crianças 0–1* apresentou maior incerteza e baixa significância, sendo parcialmente penalizado.

O Modelo Aditivo Generalizado com função de ligação logito permite incorporar relações potencialmente não lineares entre as covariáveis e a probabilidade de infecção (*pos01*). Diferentemente do modelo logístico clássico, o GAM-LR inclui funções suaves que flexibilizam essas relações, evitando a imposição de linearidade estrita e proporcionando maior capacidade de modelagem.

As funções suaves $s(\cdot)$ são estimadas por splines cúbicos com base reduzida ($k = 3$) e penalização de rugosidade, mecanismo que controla a complexidade do ajuste e evita sobreajuste (*overfitting*).

As variáveis categóricas e lineares mantêm interpretação semelhante ao modelo logístico tradicional. A Tabela 6 resume os coeficientes paramétricos estimados.

Tabela 6 – Coeficientes paramétricos do modelo GAM-LR

Variável	**Estimativa**	**Erro Padrão**	**Valor Z**	**p-valor**
Intercepto	-3.13	0.46	-6.74	1.6×10^{-11}
Histórico positivo	2.68	0.43	6.26	3.8×10^{-10}
Tip. Quilombola	1.36	0.34	4.00	6.5×10^{-05}
Tip. Ribeirinho	-0.01	1.13	-0.01	0.990
Mosquiteiro	-0.33	0.28	-1.18	0.237
Moradores/quarto	0.10	0.28	0.36	0.717
Parede risco médio/alto	0.60	0.29	2.04	0.041
Água caixa/amianto	0.67	0.38	1.78	0.076

Fonte: Elaborado pela autora.

Os efeitos não lineares são apresentados na Tabela 7.

Tabela 7 – Termos suaves do modelo GAM-LR

Termo suave	**edf**	**Ref.df**	**Chi-sq**	**p-valor**
s(quantidade_porcos_confinados)	0.00002	2	0.0000040	0.6907
s(quantidade_porcos_soltos)	0.00001	2	0.0000042	0.6705
s(quantidade_curral)	0.00001	2	0.0000007	0.9373
s(quantidade_galinhas_galinheiro)	0.00001	2	0.0000015	0.8111
s(quantidade_galinhas_soltas)	0.60500	2	2.0517030	0.0589
s(quantidade_animal_de_estimacao)	0.46659	2	1.0112290	0.1299
s(qtd_0a1)	1.80984	2	4.1931530	0.0972
s(qtd_2a12)	0.51766	2	1.0547850	0.1475
s(qtd_13a18)	0.59074	2	1.4453040	0.1133
s(qtd_19a59)	0.00001	2	0.0000053	0.5447
s(qtd_60mais)	0.69192	2	2.9637420	0.0369
s(qtd_residentes)	0.00001	2	0.0000009	0.8418
s(qtd_homens)	0.00001	2	0.0000009	0.8395
s(qtd_mulheres)	0.00001	2	0.0000036	0.4302
s(total_comodos)	0.21087	2	0.2779207	0.2434
s(relacao_residentes_quartos)	0.64730	1	2.6176460	0.0411
s(idade_responsavel)	0.00002	2	0.0000177	0.3742

Fonte: Elaborado pela autora.

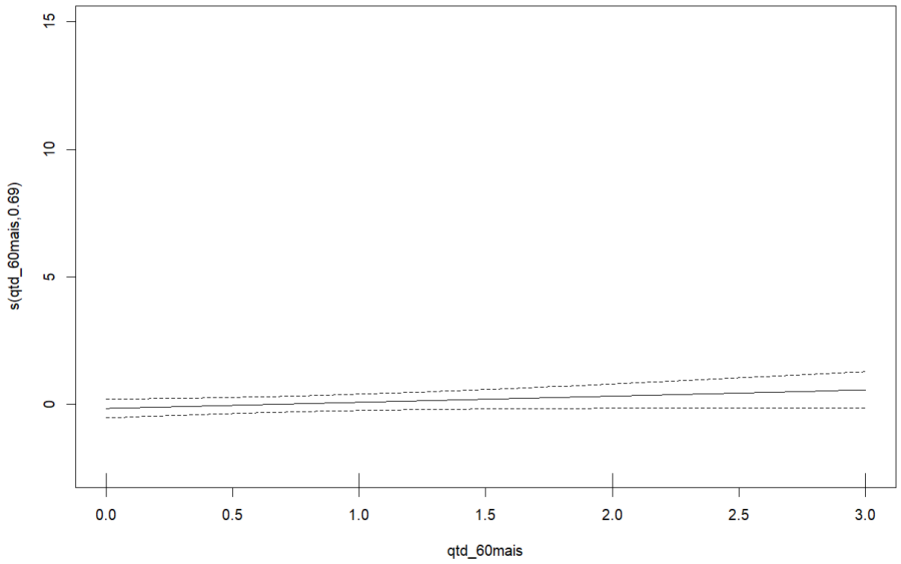
A partir dos termos suaves estimados no modelo GAM-LR, observou-se que apenas duas variáveis apresentaram evidências estatísticas robustas de não linearidade associada à presença de triatomíneos: s(qtd_60mais)p = 0,0369 e s(relacao_residentes_quartos) p = 0,0411. Esses resultados indicam que tanto a proporção de moradores com 60 anos ou mais quanto a relação de residentes por quarto exibem padrões de efeito que não são estritamente lineares, sendo mais adequadamente representados por funções suaves.

No caso de s(qtd_60mais), o termo sugere que o número de idosos no domicílio pode influenciar o risco de infestação de maneira não proporcional, possivelmente refletindo aspectos estruturais, comportamentais ou contextuais desses domicílios. Já o termo s(relacao_residentes_quartos) apresenta uma curvatura mais marcada, indicando que diferentes níveis de adensamento domiciliar podem modificar o risco de forma variável, com pontos de aumento ou estabilização ao longo da função suave.

A Figura 24 e a Figura 25 apresentam as funções suaves estimadas para as duas variáveis significativas. Esses gráficos permitem visualizar a forma da relação não linear ao longo de cada variável, destacando regiões onde o efeito estimado aumenta, se estabiliza ou diminui. No caso de quantidade com mais de 60 anos, observam-se oscilações discretas no efeito, com incrementos em determinados intervalos. Para a relação residentes quartos, a curva evidencia que o maior adensamento inicial tende a elevar a probabilidade prevista de presença de triatomíneos, seguido por zonas de estabilização do risco, reforçando a utilidade da modelagem aditiva para captar padrões complexos que modelos lineares

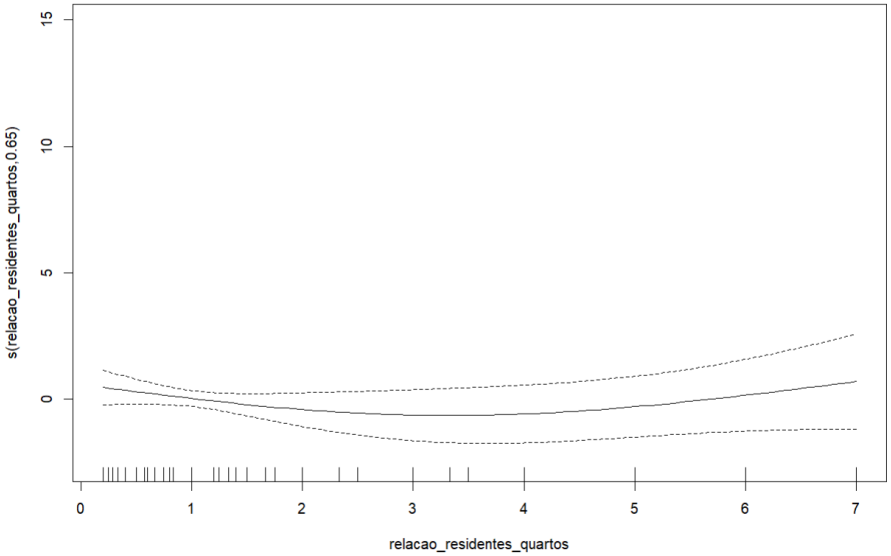
tradicionais não identificariam.

Figura 24 – Função suave estimada para `s(qtd_60mais)` no modelo GAM-LR



Fonte: Elaborado pela autora.

Figura 25 – Função suave estimada para `s(relacao_residentes_quartos)` no modelo GAM-LR



Fonte: Elaborado pela autora.

3.6 Performance do Modelo

3.6.1 Avaliação Geral

A avaliação de performance dos modelos GLM-LR e GAM-LR foi realizada a partir de múltiplas métricas clássicas em modelos de classificação binária: acurácia, sensibilidade,

especificidade, acurácia balanceada, coeficiente de correlação de Matthews (MCC) e ponto de corte (cutoff) ótimo estimado com base na maximização do MCC. Adicionalmente, foram calculadas matrizes de confusão, curvas MCC–cutoff e validação cruzada leave-one-out (LOOCV).

A Tabela 8 resume as principais métricas de desempenho obtidas.

Tabela 8 – Métricas de desempenho dos modelos GLM-LR e GAM-LR.

Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**	**Acc. Balanceada**	**MCC**	**Cutoff**
GLM	0.8588	0.3068	0.9759	0.6414	0.4115	0.45
GAM	0.8688	0.4545	0.9566	0.7056	0.4891	0.34

Fonte: Elaborado pela autora.

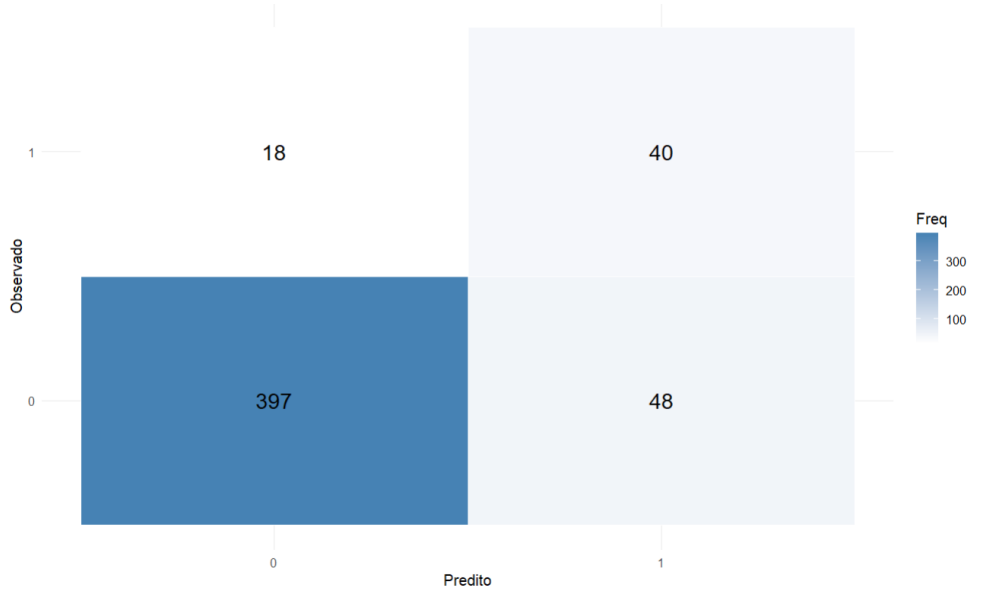
Ambos os modelos GLM-LR e GAM-LR apresentaram desempenho semelhante, com acurácia aproximada de 0,86 e 0,87, respectivamente, embora acompanhada de sensibilidade moderada, o que é esperado em cenários epidemiológicos com forte desbalanceamento entre classes. A especificidade permaneceu muito alta em ambos os casos acima de 0,95, indicando excelente capacidade de identificar corretamente os domicílios negativos para a presença de triatomíneos.

O coeficiente de Matthews (MCC) reforça essa interpretação: enquanto o GLM-LR apresentou MCC de 0,41, o GAM-LR alcançou 0,49, representando melhora consistente na correlação global entre predições e valores observados especialmente relevante em problemas desbalanceados. Além disso, a acurácia balanceada foi superior no GAM-LR 0.7056, evidenciando melhor desempenho na distinção entre positivos e negativos. Assim, embora ambos os modelos sejam adequados, o GAM-LR demonstra desempenho relativamente superior, beneficiando-se da flexibilidade dos termos suaves na captura de efeitos não lineares.

3.6.2 Matrizes de Confusão

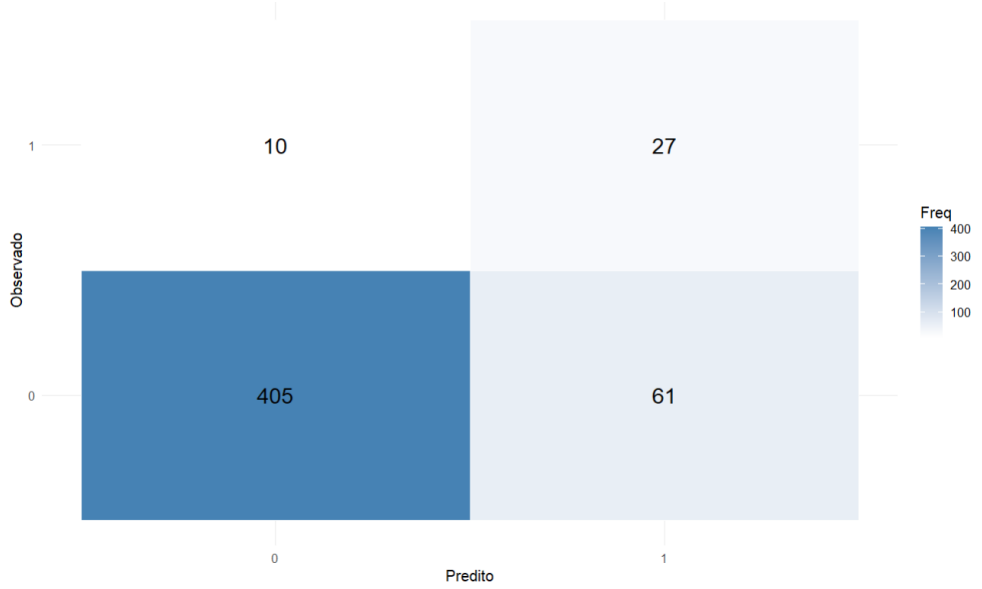
As Figuras 26 e 27 apresentam as matrizes de confusão para cada modelo, considerando o cutoff ótimo calculado com base no MCC. De forma visual, observa-se uma predominância de verdadeiros negativos e uma maior dificuldade dos modelos em identificar corretamente os verdadeiros positivos, característica consistente com bases de dados desbalanceadas.

Figura 26 – Matriz de confusão do modelo GLM-LR com cutoff = 0.45.



Fonte: Elaborado pela autora

Figura 27 – Matriz de confusão do modelo GAM-LR com cutoff = 0.34.

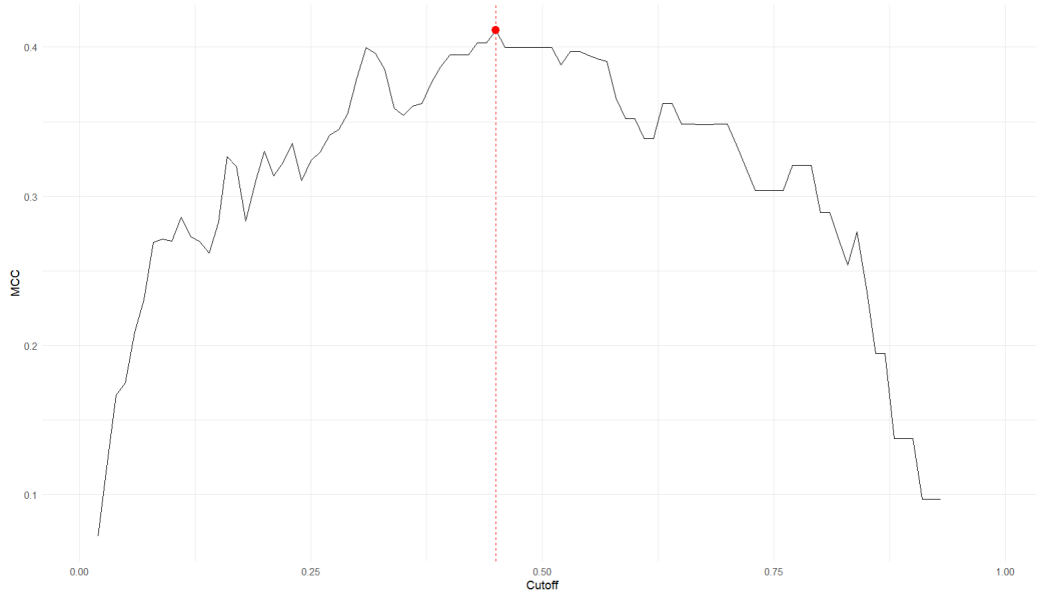


Fonte: Elaborado pela autora

3.6.3 Curva MCC–Cutoff

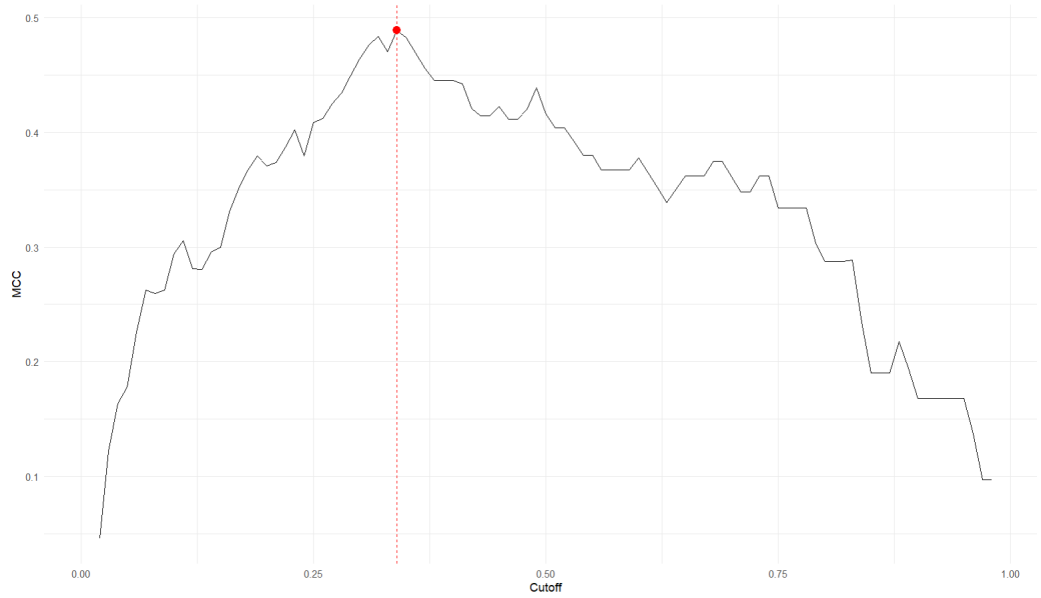
A curva MCC–cutoff permite identificar o ponto de corte que maximiza o desempenho global do modelo. Em ambos os casos, o ponto ideal situou-se entre 0.45 e 0.34, indicando estabilidade na escolha do limiar de classificação e coerência entre os diferentes ajustes.

Figura 28 – Curva MCC vs. Cutoff para o modelo GLM-LR.



Fonte: Elaborado pela autora

Figura 29 – Curva MCC vs. Cutoff para o modelo GAM-LR.



Fonte: Elaborado pela autora

3.6.4 Validação Cruzada Leave-One-Out (LOOCV)

A estabilidade preditiva e a capacidade de generalização dos modelos foram avaliadas por meio da validação cruzada do tipo *leave-one-out* (LOOCV). Nessa abordagem, cada observação é removida individualmente, o modelo é reestimado e a predição fora da amostra é registrada. Embora computacionalmente intensiva, a LOOCV oferece uma medida sólida de desempenho, especialmente útil em conjuntos de dados pequenos ou moderados.

As métricas analisadas incluíram a área sob a Curva ROC (AUC) e o coeficiente de Matthews (MCC), sendo este último particularmente indicado em cenários com possível desbalanceamento entre classes. Os resultados mostraram desempenhos muito próximos entre os modelos: o GLM-LR apresentou AUC de 0.7343, enquanto o GAM-LR registrou 0.6996. Quanto ao MCC, o GLM-LR atingiu 0.3999, com ponto de corte ótimo igual a 0.47; o GAM-LR alcançou MCC de 0.3764, com cutoff ideal igual a 0.36. Apesar da proximidade dos valores, observa-se uma leve vantagem do GLM-LR em termos puramente preditivos.

A análise conjunta das métricas indica que ambos os modelos obtiveram boa acurácia e especificidade, ainda que com sensibilidade moderada resultado esperado em bases com distribuição desbalanceada. Apesar da pequena diferença numérica, o GAM-LR apresenta maior flexibilidade estrutural, permitindo modelar relações não lineares de forma mais eficiente, o que traz ganhos interpretativos no contexto epidemiológico. Os resultados da LOOCV reforçam essa estabilidade, mostrando que ambos os modelos permanecem consistentes frente a pequenas variações nos dados. Assim, ainda que os dois sejam adequados aos objetivos propostos, o GAM-LR se destaca como uma alternativa mais robusta para análises epidemiológicas e para aplicações futuras.

4 Conclusão

Esse estudo investigou os fatores associados à presença de triatomíneos em comunidades rurais do estado de Goiás, considerando aspectos estruturais, demográficos e socioeconômicos. Por meio da aplicação de modelos de regressão logística (GLM-LR), modelos aditivos generalizados de regressão logística (GAM-LR) foi possível identificar determinantes significativos e caracterizar padrões de risco para a infestação domiciliar, oferecendo insights valiosos para a vigilância epidemiológica da doença de Chagas.

Os resultados mostram que o histórico prévio de infestação é o fator mais impactante, aumentando de forma substancial a probabilidade de presença de triatomíneos. Esse achado reforça a importância do monitoramento contínuo e da realização de intervenções direcionadas em residências previamente afetadas. Observou-se também que residências em comunidades Quilombolas apresentam risco significativamente maior de infestação, evidenciando a influência de características culturais, habitacionais e socioeconômicas na presença do vetor. Além disso, famílias com crianças menores de um ano ou com responsáveis de idade mais avançada demonstraram maior risco, indicando a necessidade de atenção especial a esses grupos.

Os aspectos estruturais das moradias, como paredes de médio ou alto risco e o uso de água proveniente de caixas de amianto, mostraram-se determinantes importantes, sugerindo que melhorias habitacionais podem contribuir para a redução da infestação. Em termos de modelagem, o GAM-LR apresentou vantagem sobre o GLM-LR.

Este trabalho contribui para a compreensão da epidemiologia da doença de Chagas ao confirmar a relevância de fatores estruturais, demográficos e socioeconômicos na infestação domiciliar por triatomíneos. Além disso, demonstra a aplicabilidade de modelos estatísticos avançados, como o GAM-LR, em contextos de vigilância epidemiológica, permitindo identificar padrões complexos que não seriam capturados por modelos lineares tradicionais. Esses achados podem apoiar a priorização de domicílios para inspeção e controle vetorial, auxiliando na formulação de estratégias de prevenção mais eficazes.

Os resultados do estudo apontam para a necessidade de programas de vigilância contínua, especialmente em residências com histórico de infestação. Intervenções habitacionais e educativas devem ser adaptadas à realidade de comunidades Quilombolas e outras populações tradicionais, garantindo a proteção de famílias mais vulneráveis. O monitoramento direcionado a famílias com crianças pequenas é fundamental, pois reduz a exposição precoce ao vetor. Ademais, a utilização de modelos preditivos pode ser incorporada como ferramenta de apoio para alocação eficiente de recursos em campanhas de controle vetorial.

Apesar da robustez metodológica, o estudo apresenta limitações. A sensibilidade dos modelos na detecção de todos os domicílios infestados foi relativamente baixa, e o pequeno tamanho de algumas categorias, como ribeirinhos, pode reduzir a precisão das estimativas. Além disso, por se tratar de um estudo transversal, não é possível estabelecer relações causais definitivas.

Para pesquisas futuras, recomenda-se a inclusão de variáveis ambientais e comportamentais mais detalhadas, a realização de estudos longitudinais para acompanhar mudanças ao longo do tempo e a exploração de técnicas de modelagem espacial e integração de diferentes fontes de dados epidemiológicos, permitindo previsões de risco mais amplas e precisas.

Em síntese, a presença de triatomíneos em comunidades rurais é um fenômeno multifatorial, influenciado por determinantes socioeconômicos, estruturais e demográficos. Este estudo evidencia que a aplicação de modelagem estatística avançada pode apoiar decisões em saúde pública, proporcionando informações valiosas para direcionar intervenções de forma mais eficiente. A integração de dados epidemiológicos com métodos preditivos representa um avanço na prevenção da doença de Chagas, contribuindo para reduzir o risco de transmissão e melhorar a qualidade de vida das comunidades afetadas.

5 REFERÊNCIAS

BATISTA, L. E.; LIMA, L. F. C. Saúde e trabalho: uma análise do perfil de morbidade segundo a ocupação no Brasil. Revista Brasileira de Saúde Ocupacional, São Paulo, v. 34, n. 119, p. 6-14, 2009.

BRASIL. Ministério da Saúde. Sistema de Informação de Agravos de Notificação – SINAN. Brasília, DF: Ministério da Saúde, 2020.

CHAO, C.; LEONE, J. L.; VIGLIANO, C. A. Chagas disease: Historic perspective. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, v. 1866, n. 5, 2020. DOI: https://doi.org/10.1016/j.bbadis.2019.165561.

DALE, C.; PASCHOALETTO, L.; COSTA, J. O Complexo Triatoma brasiliensis: atualizações sobre o principal vetor da doença de Chagas no nordeste brasileiro. Rio de Janeiro: Fundação Oswaldo Cruz, 2019.

DIAS, J. C. P. O Trypanosoma cruzi e suas características bio-ecológicas, como agente de enfermidades transmitidas por alimentos. In: Informe de la Consulta Técnica en Epidemiologia, Prevención y Manejo de la Transmisión de la Enfermedade de Chagas como Enfermedad Transmitida por Alimentos (ETA). Rio de Janeiro, RJ: OMS, 2006.

DOBSON, Annette J.; BARNETT, Adrian. An Introduction to Generalized Linear Models. 4. ed. Boca Raton: Chapman and Hall/CRC, 2018.

DORN, P. L.; MONROY, M. C.; STEVENS, L. Sustainable, integrated control of native vectors: The case of Chagas disease in Central America. Frontiers in Tropical Diseases, v. 3, 2022. DOI: https://doi.org/10.3389/fitd.2022.1000167.

DOS SANTOS, J. P. et al. Assessing the entomo-epidemiological situation of Chagas disease in rural communities in the state of Piauí, Brazilian semi-arid region. Transactions of the Royal Society of Tropical Medicine and Hygiene, v. 114, n. 11, p. 820-829, 2020. DOI: https://doi.org/10.1093/trstmh/traa070.

JURBERG, J. et al. Atlas iconográfico dos triatomíneos do Brasil: vetores da doença de Chagas. Rio de Janeiro: Instituto Oswaldo Cruz, 2014.

HASTIE, Trevor; TIBSHIRANI, Robert. Generalized Additive Models. Statistical Science, v. 1, n. 3, p. 297–318, 1986.

KROPF, S. P. Carlos Chagas e os debates e controvérsias sobre a doença no Brasil

(1909-1923). História, Ciências, Saúde – Manguinhos, v. 16, n. 1, p. 205-227, 2009. DOI: https://doi.org/10.1590/S0104-59702009000500010.

WOOD, Simon N. Generalized Additive Models: An Introduction with R. 2. ed. Boca Raton: Chapman and Hall/CRC, 2017.

WORLD HEALTH ORGANIZATION (WHO). Chagas disease (American trypanosomiasis). Geneva: WHO, 2025. Disponível em: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis). Acesso em: 28 nov. 2025.

A Apêndice

A.1 Script R Completo

```
# SCRIPT FINAL - TCC
pkgs <- c("readxl","dplyr","ggplot2","tidyr","reshape2","scales",
          "mgcv","caret","broom","pROC","MASS")
for(p in pkgs){
  if(!requireNamespace(p, quietly = TRUE)) install.packages(p)
  library(p, character.only = TRUE)
}
select <- dplyr::select
filter <- dplyr::filter
across <- dplyr::across
all_of <- rlang::syms

set.seed(123)
dados <- readxl::read_excel("C:/Users/Lalinda/Desktop/dados_tcruzi.xlsx", skip = 2)

# Criar desfecho pos01 (0/1)
if(!"pos01" %in% names(dados)){
  if("presenca_triatomineo" %in% names(dados)){
    dados$pos01 <- ifelse(dados$presenca_triatomineo == "Sim", 1, 0)
  } else stop("Crie a varivel 'presenca_triatomineo' ou informe como se chama no
    dataset.")
} else dados$pos01 <- as.integer(dados$pos01)

# Varivel mosquiteiro
if(!"mosquiteiro" %in% names(dados)){
  if("usa_protecao_de_mosquitos" %in% names(dados)){
    dados$mosquiteiro <- factor(dados$usa_protecao_de_mosquitos)
  } else if("repelente_corporal" %in% names(dados)){
    dados$mosquiteiro <- factor(dados$repelente_corporal)
    warning("Criado 'mosquiteiro' a partir de 'repelente_corporal'.")
  } else stop("Crie a varivel 'mosquiteiro' no dataset.")
} else dados$mosquiteiro <- factor(dados$mosquiteiro)

if("qtd_positivo_dc_previo" %in% names(dados)){
  tmp_qtd <- suppressWarnings(as.numeric(as.character(dados$qtd_positivo_dc_previo)))
  dados$qtd_positivo_dc_previo <- ifelse(is.na(tmp_qtd), NA, ifelse(tmp_qtd > 0, 1, 0))
  dados$qtd_positivo_dc_previo <- factor(dados$qtd_positivo_dc_previo, levels = c(0,1),
                                         labels = c("No","Sim"))
} else {
  warning("Varivel 'qtd_positivo_dc_previo' no encontrada no dataset  verifique nomes.")
}

# Variveis  contnuas
num_vars <- c(
  "quantidade_porcos_confinados","quantidade_porcos_soltos","quantidade_curral",
    "quantidade_galinhas_galinheiro","quantidade_galinhas_soltas","quantidade_animal_de_estimao",
  "qtd_0a1","qtd_2a12","qtd_13a18","qtd_19a59","qtd_60mais",
```

```
47
    "qtd_residentes","qtd_homens","qtd_mulheres","total_comodos","relacao_residentes_quartos",
48  "idade_responsavel"
49  )
50  num_vars <- intersect(num_vars, names(dados))
51  dados[num_vars] <- lapply(dados[num_vars], function(x) as.numeric(as.character(x)))
52
53  # Variveis  categricas
54  fac_vars <- c("comu_cod_nom","cod_mun","tipologia_comunidade",
55                "parede_medio_alto_risco","agua_lavar_alimentos_caixa_d_agua_amianto",
56                "presenca_triatomineo","mosquiteiro","qtd_positivo_dc_previo")
57  fac_vars <- intersect(fac_vars, names(dados))
58  dados[fac_vars] <- lapply(dados[fac_vars], factor)
59
60  # Tratar infinitos e limpar
61  dados <- dados %>% mutate(across(where(is.numeric), ~ ifelse(is.infinite(.), NA_real_,
    .)))
62  dados_clean <- dados %>% filter(!is.na(pos01))
63
64  # Contnuas
65  if(length(num_vars) > 0){
66    desc_cont <- dados_clean %>%
67      dplyr::select(all_of(num_vars)) %>%
68      summarise(across(everything(), list(
69      mean = ~mean(. , na.rm=TRUE),
70      median = ~median(. , na.rm=TRUE),
71      sd = ~sd(. , na.rm=TRUE),
72      min = ~min(. , na.rm=TRUE),
73      max = ~max(. , na.rm=TRUE)
74      ), .names="{.col}_{.fn}"))
75    cat("\n=== Estatstica Descritiva: Contnuas ===\n")
76    print(desc_cont)
77  }
78
79  # Categricas
80  if(length(fac_vars) > 0){
81    cat("\n=== Estatstica Descritiva: Categricas ===\n")
82    tab_cat <- lapply(fac_vars, function(var){
83      if(var %in% names(dados_clean)){
84        tab <- table(dados_clean[[var]], useNA="ifany")
85        prop_tab <- prop.table(tab)
86        data.frame(
87          Varivel = var,
88          Nvel  = names(tab),
89          Freq = as.vector(tab),
90          Perc = round(as.vector(prop_tab)*100,2)
91        )
92      }
93    })
94    tab_cat <- do.call(rbind, tab_cat)
95    print(tab_cat)
96  }
97
98  ## Estatstica descritiva
99  cat("=== Estatstica Descritiva: Contnuas ===\n")
100 if(length(num_vars) > 0){
101   desc_cont <- dados_clean %>%
```

```
    dplyr::select(all_of(num_vars)) %>%
    summarise(across(everything(), list(
    mean = ~mean(. , na.rm=TRUE),
    median = ~median(. , na.rm=TRUE),
    sd = ~sd(. , na.rm=TRUE),
    min = ~min(. , na.rm=TRUE),
    max = ~max(. , na.rm=TRUE)
    ), .names="{.col}_{.fn}"))
  print(desc_cont)
} else cat("Nenhuma varivel contnua encontrada.\n")

out_dir <- "graficos_categoricos"
if(!dir.exists(out_dir)) dir.create(out_dir)

# Variveis alvo para grficos
fac_vars_graficos <- intersect(
  c("cod_mun",
    "tipologia_comunidade",
    "parede_medio_alto_risco",
    "agua_lavar_alimentos_caixa_d_agua_amianto",
    "presenca_triatomineo",
    "mosquiteiro",
    "qtd_positivo_dc_previo"),
  names(dados_clean)
)

# Estatstica descritiva categrica
for(v in fac_vars_graficos){
  df_plot <- dados_clean %>%
    dplyr::select(all_of(c(v,"presenca_triatomineo"))) %>%
    filter(!is.na(.data[[v]]), !is.na(presenca_triatomineo)) %>%
    mutate(var = factor(.data[[v]]),
           Triatomineo = factor(presenca_triatomineo, levels = c("No","Sim"),
                                labels = c("No","Sim")))

  # Contagem
  p1 <- ggplot(df_plot, aes(x = var, fill = Triatomineo)) +
    geom_bar() +
    labs(title=paste("Barras -", v), x=v, y="N de Domiclios", fill="Triatomneos") +
    theme_minimal(base_size=12) + theme(axis.text.x = element_text(angle=45, hjust=1))
  ggsave(file.path(out_dir, paste0("empilhada_contagem_", v, ".png")), p1, width=8,
    height=5, dpi=300)

  #  Proporo
  df_sum <- df_plot %>% count(var, Triatomineo) %>% group_by(var) %>% mutate(prop = n /
    sum(n))
  p2 <- ggplot(df_sum, aes(x = var, y = prop, fill = Triatomineo)) +
    geom_col(position="stack") +
    scale_y_continuous(labels = scales::percent) +
    labs(title=paste("Proporo (%) -", v), x=v, y="N de Domiclios", fill="Triatomneos") +
    theme_minimal(base_size=12) + theme(axis.text.x = element_text(angle=45, hjust=1))
  ggsave(file.path(out_dir, paste0("empilhada_proporcao_", v, ".png")), p2, width=8,
    height=5, dpi=300)

  message("Grficos gerados: ", v)
}
```

```
# Boxplot violino para variveis contnuas
dados_clean <- dados_clean %>%
  mutate(pos01_factor = factor(pos01, levels=c(0,1), labels=c("0 - Ausncia","1 -
    Presena")))
for(v in num_vars){
  if(v %in% names(dados_clean)){
    df_v <- dados_clean %>% dplyr::select(pos01_factor, all_of(v)) %>% rename(valor =
    !!v) %>% filter(!is.na(valor))
    p <- ggplot(df_v, aes(x = pos01_factor, y = valor, fill = pos01_factor)) +
      geom_violin(alpha = 0.4, trim = TRUE) +
      geom_boxplot(width = 0.15, fill = "white") +
      scale_fill_manual(values = c("0 - Ausncia"="violet","1 - Presena"="purple")) +
      labs(title=paste("Distribuio:", v), x="pos01", y=v) + theme_minimal(base_size=12)
    print(p)
  }
}

## Testes de diferena
cat("\n=== Testes de diferena entre grupos (contnuas) ===\n")
for(v in num_vars){
  if(v %in% names(dados_clean)){
    group0 <- dados_clean[[v]][dados_clean$pos01==0]
    group1 <- dados_clean[[v]][dados_clean$pos01==1]
    n0 <- length(na.omit(group0)); n1 <- length(na.omit(group1))
    cat("\nVarivel:", v, "\n")
    if(n0 < 3 | n1 < 3){
      cat("  Grupos pequenos -> Wilcoxon\n")
      print(wilcox.test(group0, group1))
      next
    }
    p0 <- tryCatch(shapiro.test(na.omit(group0))$p.value, error=function(e) NA)
    p1 <- tryCatch(shapiro.test(na.omit(group1))$p.value, error=function(e) NA)
    cat("Shapiro p (grupo0):", round(p0,4), " Shapiro p (grupo1):", round(p1,4), "\n")
    if(!is.na(p0) && !is.na(p1) && p0 > 0.05 && p1 > 0.05){
      cat("  Normalidade plausvel -> t.test (Welch)\n"); print(t.test(group0, group1,
    var.equal = FALSE))
    } else {
      cat("   No -normal -> Wilcoxon\n"); print(wilcox.test(group0, group1))
    }
  }
}

## Modelos: GLM (logstico) e GAM (logstico)
form_glm <- pos01 ~ tipologia_comunidade + idade_responsavel +
  qtd_0a1 + qtd_positivo_dc_previo + mosquiteiro +
  relacao_residentes_quartos + parede_medio_alto_risco +
  agua_lavar_alimentos_caixa_d_agua_amianto

smooth_vars <- intersect(num_vars, names(dados_clean))

s_terms <- paste0("s(", smooth_vars, ", k=3)")

# variveis paramtricas (categricas)
param_vars <- c("qtd_positivo_dc_previo","tipologia_comunidade","mosquiteiro",
                "relacao_residentes_quartos","parede_medio_alto_risco",
                "agua_lavar_alimentos_caixa_d_agua_amianto")
param_vars <- intersect(param_vars, names(dados_clean))
```

```
210
211 form_gam_str <- paste("pos01 ~", paste(c(s_terms, param_vars), collapse = " + "))
212 form_gam <- as.formula(form_gam_str)
213 cat("\nForm GAM gerada:\n"); print(form_gam)
214
215 # criar dados completos para os modelos
216 vars_model <- unique(all.vars(form_glm))
217 linhas <- complete.cases(dados_clean[, vars_model])
218 dados_clean_complete <- dados_clean[linhas, ]
219
220 # Estimar modelos em dados completos
221 mod_glm <- glm(form_glm, data = dados_clean_complete, family = binomial)
222 mod_gam <- mgcv::gam(form_gam, data = dados_clean_complete, family = binomial, select =
        TRUE)
223
224 # Previses (probabilidades)
225 prob_glm <- predict(mod_glm, type="response")
226 prob_gam <- predict(mod_gam, type="response")
227
228 # Resultados GLM - outputs principais
229 cat("\n=== Resultados GLM ===\n")
230 print(summary(mod_glm))
231 vcov_glm <- vcov(mod_glm)
232 se_glm <- sqrt(diag(vcov_glm))
233 wald_table <- broom::tidy(mod_glm)
234 cat("\n=== Wald tests (GLM) ===\n"); print(wald_table)
235
236 # Odds ratios e IC 95%
237 or_glm <- exp(coef(mod_glm))
238 ci_glm <- exp(confint.default(mod_glm))
239 or_table_glm <- data.frame(term = names(or_glm), OR = or_glm, CI_low = ci_glm[,1],
        CI_high = ci_glm[,2])
240 cat("\n=== Odds Ratios (GLM) ===\n"); print(or_table_glm)
241
242 # Resultados GAM - outputs principais
243 cat("\n=== Resultados GAM ===\n")
244 sum_gam <- summary(mod_gam)   # guardar summary para inspeo
245 print(sum_gam)
246 cat("\n--- Tabela de termos suaves (s.table) ---\n")
247 print(sum_gam$s.table)
248
249 cat("\n--- Coeficientes paramtricos ---\n")
250 print(sum_gam$p.table)
251
252 #Plotar todos os suavizados significativos
253 out_dir_smooth <- "graficos_suavizados"
254 if(!dir.exists(out_dir_smooth)) dir.create(out_dir_smooth)
255
256 # sum_gam$s.table rows tm nomes como "s(idade_responsavel)"
257 if(!is.null(sum_gam$s.table)){
258   s_table <- as.data.frame(sum_gam$s.table)
259   s_table$term <- rownames(s_table)
260   # coluna de p-value pode ter nome "p-value" ou "p.value" dependendo da verso; buscar
261   pcol <- grep("p", tolower(names(s_table)), value = TRUE)
262   # preferir coluna que contenha "p"
263   pcol <- pcol[1]
264   s_table$pv <- s_table[[pcol]]
```

```
  signif_smooths <- s_table %>% filter(!is.na(pv) & pv < 0.05)
  if(nrow(signif_smooths) == 0){
    message("Nenhum termo suavizado com p < 0.05 encontrado.")
  } else {
    message("Suavizados significativos encontrados:\n")
    print(signif_smooths[, c("term","pv")])
    # Para cada termo significativo, gerar plot (usando plot.gam)
    for(i in seq_len(nrow(s_table))){
      term_name <- s_table$term[i]
      # se esse termo for significativo, plotar e salvar
      if(term_name %in% signif_smooths$term){
        png(filename = file.path(out_dir_smooth, paste0(gsub("[^A-Za-z0-9_]", "_",
    term_name), ".png")),
            width = 800, height = 600)
        # plot.gam usa ndice do smooth: encontrar ndice correspondente
        # match by term name in the GAM object's smooth labels
        smooth_labels <- sapply(mod_gam$smooth, function(x) x$label)
        sel_idx <- which(smooth_labels == term_name)
        if(length(sel_idx) == 1){
          plot(mod_gam, select = sel_idx, shade = TRUE, seWithMean = TRUE, main =
    term_name)
        } else {
          # fallback: plot all and hope graphic shows it (rare)
          plot(mod_gam, shade = TRUE, seWithMean = TRUE, pages = 1)
        }
        dev.off()
        message("Grfico salvo:", file.path(out_dir_smooth, paste0(gsub("[^A-Za-z0-9_]",
    "_", term_name), ".png")))
        # tambm mostrar na sesso
        if(length(sel_idx) == 1){
          plot(mod_gam, select = sel_idx, shade = TRUE, seWithMean = TRUE, main =
    term_name)
        }
      }
    }
  }
} else {
  warning("Modelo GAM no retornou tabela de termos suaves (s.table).")
}

### Seleo de variveis e comparao (StepAIC/AIC/BIC e select=TRUE)
#Aplicar stepAIC (GLM)
mod_glm_full <- glm(form_glm, data = dados_clean_complete, family = binomial)
mod_glm_step <- MASS::stepAIC(mod_glm_full, direction = "both", trace = FALSE)
cat("\n=== GLM Stepwise (stepAIC) resumo ===\n")
print(summary(mod_glm_step))
cat("AIC (step model):", AIC(mod_glm_step), "\n")
cat("BIC (step model):", BIC(mod_glm_step), "\n")

#Aplicar GAM com select=TRUE
cat("\n=== GAM (select=TRUE) resumo ===\n")
print(summary(mod_gam)) # mostra quais termos receberam penalizao

# Comparar AIC/BIC
cat("\ nComparao  AIC/BIC (GLM full vs GLM step):\n")
cat("AIC full:", AIC(mod_glm_full), " AIC step:", AIC(mod_glm_step), "\n")
cat("BIC full:", BIC(mod_glm_full), " BIC step:", BIC(mod_glm_step), "\n")
```

```
318
319 # Comparar GLM vs GAM, usar AIC
320 AIC_gam <- AIC(mod_gam)
321 cat("AIC GAM:", AIC_gam, "\n")
322
323 ## MTRICAS DE DESEMPENHO
324
325 #  Funo  MCC robusta
326 calc_mcc_vec <- function(y_true, y_pred){
327
328   stopifnot(length(y_true) == length(y_pred))
329   y_true <- as.numeric(y_true)
330   y_pred <- as.numeric(y_pred)
331   valid <- !is.na(y_true) & !is.na(y_pred)
332   if(sum(valid) == 0) return(NA_real_)
333
334   TP <- sum(y_true[valid]==1 & y_pred[valid]==1)
335   TN <- sum(y_true[valid]==0 & y_pred[valid]==0)
336   FP <- sum(y_true[valid]==0 & y_pred[valid]==1)
337   FN <- sum(y_true[valid]==1 & y_pred[valid]==0)
338
339   TP <- as.numeric(TP); TN <- as.numeric(TN); FP <- as.numeric(FP); FN <- as.numeric(FN)
340
341   den <- sqrt( (TP+FP) * (TP+FN) * (TN+FP) * (TN+FN) )
342   if(is.na(den) || den == 0) return(NA_real_)
343   (TP*TN - FP*FN) / den
344 }
345
346 #  Funo  cutoff timo via MCC (tratando NA)
347 cutoff_mcc <- function(probs, labels, by=0.01){
348   probs <- as.numeric(probs)
349   labels <- as.numeric(labels)
350   thr_seq <- seq(0.01, 0.99, by=by)
351   mccs <- sapply(thr_seq, function(th){
352   preds <- as.integer(probs >= th)
353     calc_mcc_vec(labels, preds)
354   })
355   if(all(is.na(mccs))){
356     return(list(cutoff = NA_real_, max_mcc = NA_real_, df = data.frame(Cutoff=thr_seq,
      MCC=mccs)))
357   }
358   mccs_filled <- ifelse(is.na(mccs), -Inf, mccs)
359   best_idx <- which.max(mccs_filled)
360   list(cutoff = thr_seq[best_idx], max_mcc = mccs[best_idx], df =
      data.frame(Cutoff=thr_seq, MCC=mccs))
361 }
362 # Labels verdadeiros
363 y_true <- as.numeric(dados_clean_complete$pos01)  # j deve ser 0/1
364
365 # Determinar cutoffs
366 res_glm_cut <- cutoff_mcc(prob_glm, y_true)
367 res_gam_cut <- cutoff_mcc(prob_gam, y_true)
368
369 cutoff_glm <- ifelse(is.na(res_glm_cut$cutoff), 0.5, res_glm_cut$cutoff)
370 cutoff_gam <- ifelse(is.na(res_gam_cut$cutoff), 0.5, res_gam_cut$cutoff)
371
372 #  Predies  binrias
```

```
pred_glm01 <- as.integer(prob_glm >= cutoff_glm)
pred_gam01 <- as.integer(prob_gam >= cutoff_gam)

# Matriz de Confuso
y_true_f <- factor(y_true, levels = c(0,1))
pred_glm_f <- factor(pred_glm01, levels = c(0,1))
pred_gam_f <- factor(pred_gam01, levels = c(0,1))

cm_glm <- caret::confusionMatrix(pred_glm_f, y_true_f, positive="1")
cm_gam <- caret::confusionMatrix(pred_gam_f, y_true_f, positive="1")

# Extrair mtricas do caret (com fallback se NA)
safe_extract <- function(cm){
  out <- list()
  out$Accuracy <- as.numeric(cm$overall["Accuracy"])
  # byClass pode ser um vetor nomeado
  byc <- cm$byClass
  out$Sensitivity <- as.numeric(byc["Sensitivity"])
  out$Specificity <- as.numeric(byc["Specificity"])
  out$Balanced <- as.numeric(byc["Balanced Accuracy"])
  out
}
m_glm <- safe_extract(cm_glm)
m_gam <- safe_extract(cm_gam)

mcc_glm <- calc_mcc_vec(y_true, pred_glm01)
mcc_gam <- calc_mcc_vec(y_true, pred_gam01)

metrics_table <- data.frame(
  Model = c("GLM", "GAM"),
  Accuracy = c(m_glm$Accuracy, m_gam$Accuracy),
  Sensitivity = c(m_glm$Sensitivity, m_gam$Sensitivity),
  Specificity = c(m_glm$Specificity, m_gam$Specificity),
  Balanced_Accuracy = c(m_glm$Balanced, m_gam$Balanced),
  MCC = c(mcc_glm, mcc_gam),
  Cutoff = c(cutoff_glm, cutoff_gam),
  row.names = NULL,
  stringsAsFactors = FALSE
)
cat("\n=== Tabela de Mtricas ===\n")
print(metrics_table)

# Plotar matriz de confuso - GLM e GAM
plot_cm_heatmap <- function(cm_obj, title = "Matriz de Confuso"){
  cm_df <- as.data.frame(cm_obj$table)
  names(cm_df) <- c("Reference","Prediction","Freq")
  cm_df$Reference <- factor(cm_df$Reference, levels=c("0","1"))
  cm_df$Prediction <- factor(cm_df$Prediction, levels=c("0","1"))
  p <- ggplot(cm_df, aes(x=Prediction, y=Reference, fill=Freq)) +
    geom_tile(color="white") +
    geom_text(aes(label=Freq), size=6) +
    scale_fill_gradient(low="white", high="steelblue") +
    labs(title=title, x="Predito", y="Observado") +
    theme_minimal(base_size=12)
  print(p)
}
plot_cm_heatmap(cm_glm, sprintf("GLM (cutoff=%.2f)  MCC=%.3f", cutoff_glm, mcc_glm))
```

```
plot_cm_heatmap(cm_gam, sprintf("GAM (cutoff=%.2f)  MCC=%.3f", cutoff_gam, mcc_gam))

#Plot MCC vs Cutoff
df_mcc_glm <- res_glm_cut$df
df_mcc_gam <- res_gam_cut$df

if(all(is.na(df_mcc_glm$MCC))){
  warning("Todos MCCs NA para GLM  no  h  cutoffs vlidos para plotar.")
} else {
  ggplot(df_mcc_glm, aes(Cutoff, MCC)) +
    geom_line() + geom_vline(xintercept = cutoff_glm, linetype="dashed", color="red") +
    geom_point(aes(x=cutoff_glm, y=res_glm_cut$max_mcc), color="red", size=3) +
    labs(title="GLM: MCC vs Cutoff") + theme_minimal()
}

if(all(is.na(df_mcc_gam$MCC))){
  warning("Todos MCCs NA para GAM  no  h  cutoffs vlidos para plotar.")
} else {
  ggplot(df_mcc_gam, aes(Cutoff, MCC)) +
    geom_line() + geom_vline(xintercept = cutoff_gam, linetype="dashed", color="red") +
    geom_point(aes(x=cutoff_gam, y=res_gam_cut$max_mcc), color="red", size=3) +
    labs(title="GAM: MCC vs Cutoff") + theme_minimal()
}

## Odds Ratios e Forest plots (GLM)
or_glm_df <- broom::tidy(mod_glm, exponentiate = TRUE, conf.int = TRUE) %>%
  dplyr::rename(OR = estimate, CI_low = conf.low, CI_high = conf.high) %>%
  dplyr::select(term, OR, CI_low, CI_high, p.value)
print(or_glm_df)

# Forest plot GLM
or_sig <- or_glm_df %>% filter(p.value < 0.05 & term != "(Intercept)")
if(nrow(or_sig) > 0){
  p_or <- ggplot(or_sig, aes(x = reorder(term, OR), y = OR)) +
    geom_point() + geom_errorbar(aes(ymin=CI_low, ymax=CI_high), width=0.2) +
    geom_hline(yintercept=1, linetype="dashed", color="red") + coord_flip() +
    theme_minimal()
  print(p_or)
}

## LOOCV
loocv_mcc <- function(model_type = c("GLM","GAM"), formula, data){
  model_type <- match.arg(model_type)
  n <- nrow(data)
  probs <- numeric(n)
  y_true <- as.numeric(data$pos01)
  for(i in seq_len(n)){
    train <- data[-i,,drop=FALSE]
    test <- data[i,,drop=FALSE]
    if(model_type == "GLM"){
      fit <- glm(formula, data = train, family = binomial)
    } else {
      fit <- mgcv::gam(formula, data = train, family = binomial, select = TRUE)
    }
    probs[i] <- predict(fit, newdata = test, type = "response")
  }
  best <- cutoff_mcc(probs, y_true)
```

```
  preds01 <- as.integer(probs >= best$cutoff)
  mcc_val <- calc_mcc_vec(y_true, preds01)
  list(Prob = probs, Cutoff = best$cutoff, Max_MCC = best$max_mcc, MCC = mcc_val)
}

cat("\n=== Executando LOOCV (pode levar tempo) ===\n")
loocv_glm_res <- loocv_mcc("GLM", form_glm, dados_clean)
cat("LOOCV GLM - MCC:", loocv_glm_res$MCC, " Cutoff:", loocv_glm_res$Cutoff, "\n")
loocv_gam_res <- loocv_mcc("GAM", form_gam, dados_clean)
cat("LOOCV GAM - MCC:", loocv_gam_res$MCC, " Cutoff:", loocv_gam_res$Cutoff, "\n")

# LOOCV AUC
y_loocv <- as.numeric(dados_clean$pos01)
auc_glm_loocv <- pROC::roc(y_loocv, loocv_glm_res$Prob)$auc
auc_gam_loocv <- pROC::roc(y_loocv, loocv_gam_res$Prob)$auc
cat("LOOCV GLM - AUC:", auc_glm_loocv, "\n")
cat("LOOCV GAM - AUC:", auc_gam_loocv, "\n")
```

Listing A.1 – Script completo em R