Anna Clara Damasio Monteiro

Análise de Matrizes de Contato Suavizadas no Complexo de Manguinhos

Rio de Janeiro
2025

Anna Clara Damasio Monteiro

Análise de Matrizes de Contato Suavizadas no Complexo de Manguinhos

Projeto de Monografia apresentado à Escola de Matemática Aplicada como requisito parcial para continuidade ao trabalho de monografia.

Aprovado em ____ / ____ / ____

Grau atribuído ao Projeto de Monografia: _____

Prof. Dr. Claudio José Struchiner
Orientador
Escola de Matemática Aplicada
Fundação Getúlio Vargas

Rio de Janeiro
2025

Agradecimentos

Agradeço primeiramente a Deus e toda espiritualidade por guiarem meus passos, permitindo-me chegar até aqui.

À minha família, que sempre esteve ao meu lado, e em especial à minha mãe. Mesmo distante fisicamente, ela se fez presente em cada etapa, tornando esta caminhada mais leve, tranquila e possível graças ao seu amor e incentivo incondicional.

Ao meu orientador, Prof. Claudio José Struchiner, expresso minha profunda gratidão pela orientação, atenção e pela confiança depositada no desenvolvimento deste trabalho.

Aos professores da FGV EMAp, pelos ensinamentos compartilhados, que foram essenciais para a minha formação.

Por fim, deixo um agradecimento especial à Cássia e à Claudinha, que foram essenciais para a minha trajetória acadêmica dentro da FGV. Obrigada por todo o suporte, carinho e por sempre acreditarem em mim e me apoiarem durante esta jornada.

Resumo

A modelagem matemática de doenças infecciosas depende criticamente de matrizes de contato social precisas para estimar parâmetros epidemiológicos como o R_0. No entanto, matrizes genéricas frequentemente falham em capturar as dinâmicas de interação em territórios vulneráveis. Este trabalho apresenta a construção e validação de matrizes de contato suavizadas para o Complexo de Manguinhos (RJ), utilizando dados do inquérito sorológico COMVIDA. Diante de um severo sub-relato de contatos intradomiciliares e alta sobredispersão nos dados brutos, aplicou-se inicialmente uma correção metodológica baseada no mínimo domiciliar esperado. Para a suavização das taxas, comparou-se o desempenho de técnicas não-paramétricas (LOESS) e Modelos Aditivos Generalizados (GAM), avaliando diferentes famílias de distribuição (Poisson, Quasi-Poisson e Binomial Negativa) e estruturas geométricas de predição (Idade-Idade vs. Coorte). Os resultados demonstraram que o método LOESS e a família Poisson foram incapazes de lidar com a heterocedasticidade dos dados. A melhor aderência foi obtida pelo modelo GAM com distribuição Binomial Negativa e base de splines de dimensão $k = 20$ na geometria cartesiana padrão. Este modelo superou a abordagem de coorte (que apresentou *overfitting*) alcançando o menor erro de validação cruzada (CV-RMSE = 0,079) e uma Deviance Explicada de $99,8\%$. O estudo entrega uma matriz de contato inédita e estatisticamente robusta para Manguinhos, capturando padrões de assortatividade e mistura intergeracional essenciais para simulações epidemiológicas fidedignas em favelas. .[1]

Palavras-chave: Matrizes de Contato. Epidemiologia Matemática. Modelos Aditivos Generalizados (GAM). LOESS. Complexo de Manguinhos. COVID-19.

[1] O código para reproduzir os resultados do presente trabalho está disponível em: <https://github.com/damasio23/manguinhos-contact-matrices.git>

Abstract

Mathematical modeling of infectious diseases relies critically on accurate social contact matrices to estimate epidemiological parameters such as R_0. However, generic matrices often fail to capture interaction dynamics in vulnerable territories. This study presents the construction and validation of smoothed contact matrices for the Complexo de Manguinhos (RJ), using data from the COMVIDA serological survey. Facing severe under-reporting of intra-household contacts and high overdispersion in raw data, a methodological correction based on the expected household minimum was applied. For rate smoothing, the performance of non-parametric techniques (LOESS) and Generalized Additive Models (GAM) was compared, evaluating different distribution families (Poisson, Quasi-Poisson, and Negative Binomial) and geometric prediction structures (Age-Age vs. Cohort). Results showed that LOESS and Poisson families were unable to handle data heteroscedasticity. The best fit was achieved by the GAM model with a Negative Binomial distribution and a spline basis dimension of $k = 20$ in standard Cartesian geometry. This model outperformed the cohort approach—which showed overfitting and visual artifacts—achieving the lowest cross-validation error (CV-RMSE = 0.079) and an Explained Deviance of 99.8%. The study provides a novel, statistically robust contact matrix for Manguinhos, capturing assortativity and intergenerational mixing patterns essential for reliable epidemiological simulations in slums (*favelas*). [2]

Keywords: Contact Matrices. Mathematical Epidemiology. Generalized Additive Models (GAM). LOESS. Complexo de Manguinhos. COVID-19.

[2] The code to reproduce the results of this work is available at: <https://github.com/damasio23/manguinhos-contact-matrices.git>

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A modelagem matemática de doenças infecciosas tornou-se um pilar fundamental para a vigilância e resposta em saúde pública. A capacidade de antecipar a trajetória de uma epidemia e avaliar o impacto de intervenções, como o fechamento de escolas ou a vacinação, depende criticamente da precisão de seus parâmetros (VANDENDIJCK et al., 2024). No cerne desses modelos, especialmente para patógenos de transmissão respiratória ou de contato próximo, estão os padrões de mistura social, que ditam "quem transmite para quem" (MELEGARO et al., 2011).

Esses padrões são quantificados por meio de matrizes de contato. Uma matriz de contato é uma estrutura de dados que estima a frequência média diária de interações entre diferentes subgrupos populacionais, tradicionalmente estratificados por faixa etária. Em levantamentos clássicos, como o estudo europeu POLYMOD, um "contato" é rigorosamente definido para capturar interações epidemiologicamente relevantes, como o contato físico pele a pele ou uma conversa bidirecional próxima (MOSSONG et al., 2008).

A principal finalidade dessas matrizes é servir como um parâmetro empírico para modelos epidemiológicos. Elas substituem suposições genéricas sobre a mistura populacional por dados observados, alimentando diretamente o cálculo de métricas cruciais, como o número básico de reprodução (R_0) e a força da infecção, permitindo assim simulações mais realistas de cenários de transmissão (MELEGARO et al., 2011; KASSTEELE; EIJKEREN; WALLINGA, 2017).

Contudo, a dinâmica de transmissão não é homogênea e transcende a idade. Fatores estruturais e socioeconômicos moldam profundamente os padrões de interação e afetam o risco de infecção (MANNA et al., 2024). Este estudo foca no Complexo de Manguinhos, no Rio de Janeiro, um território que exemplifica essa interação. Composto por dezesseis comunidades, Manguinhos é uma das áreas mais pobres da cidade, apresentando o quinto pior Índice de Desenvolvimento Humano (IDH) do município.

Conforme caracterizado por Coelho et al. (2022), este cenário apresenta "pobreza concentrada, condições de moradia inseguras e inadequadas". O perfil dos moradores levantado pelo inquérito sorológico COMVIDA-Fiocruz reforça esta vulnerabilidade: 40,7% da população é parda e 22,7% é negra; 44% possuem menos de 10 anos de escolaridade; e 40,4% das famílias sobrevivem com renda mensal de até um salário mínimo. Adicionalmente, as condições estruturais incluem alta densidade domiciliar (69,8% dos domicílios com três ou mais moradores) e alta dependência de transporte coletivo (43,6% dos participantes).

A pandemia de COVID-19 expôs de forma trágica como as desigualdades socioeconômicas são determinantes na propagação de doenças infecciosas. Em territórios vulneráveis, a incapacidade de aderir a medidas de distanciamento social, seja pela necessidade de

trabalhar ou pelas condições de moradia, altera drasticamente os padrões de contato.

O Complexo de Manguinhos foi severamente afetado pela pandemia. Um inquérito sorológico de base populacional conduzido na área entre setembro de 2020 e fevereiro de 2021 (o "Estudo 1" do COMVIDA) encontrou uma soroprevalência de 49,0% (para anticorpos Anti-S IgG), um índice muito superior ao de outras áreas da cidade no mesmo período. Esta alta transmissão é impulsionada por condições estruturais, não por falha individual (COELHO et al., 2022).

A combinação de domicílios numerosos e multigeracionais, a dependência de transporte público lotado e a necessidade de trabalho presencial cria redes de contato densas e assimétricas. Isso amplifica o risco, fazendo com que a população local atue sistemicamente como "super-espalhadora".

No entanto, os dados brutos de contato do COMVIDA apresentam um desafio metodológico significativo: 64,38% dos entrevistados relataram *zero* contatos no dia anterior, um número incompatível com a alta densidade domiciliar (69,8% de casas com 3+ pessoas) e a alta soroprevalência (49,0%) observadas. Este sub-relato, provável omissão de contatos domiciliares, torna a matriz de taxas bruta ruidosa, esparsa e inadequada para modelagem direta. Além deste sub-relato sistemático, a matriz bruta sofre de outras fontes de ruído inerentes a levantamentos de diário: o ruído estocástico, referente a variações amostrais que geram células na matriz com pouca ou nenhuma observação; o viés de recordação, onde os participantes podem esquecer contatos menos frequentes; e um potencial viés de amostragem, dado que o próprio Estudo 1 do COMVIDA reportou uma alta taxa de falha de inclusão (67%) (COELHO et al., 2022). Portanto, a construção de uma matriz de contato específica para Manguinhos exige não apenas o uso de dados locais, mas também um tratamento estatístico robusto para corrigir o sub-relato e aplicar métodos de suavização que filtrem o ruído amostral sem distorcer os padrões únicos de interação desta população.

Modelar uma matriz de contato suavizada que reflita realisticamente os padrões de contato social ocorridos durante a pandemia de COVID-19 no Complexo de Manguinhos, utilizando dados empíricos do Estudo 1 do COMVIDA e corrigindo o sub-relato de contatos domiciliares, de modo a disponibilizar a matriz suavizada final como um parâmetro validado para futuras simulações epidemiológicas no território.

Para suavizar as taxas brutas Γ_{ij} e gerar uma matriz adequada a simulações epidemiológicas, empregamos duas técnicas não-paramétricas: LOESS (Local Estimated Scatterplot Smoothing) e GAM (Modelos Aditivos Generalizados). Ambas exploram a premissa de que Γ_{ij} varia de forma relativamente contínua com as idades (idade do respondente x_i e idade do contato y_j) e, alternativamente, pela perspectiva de coorte (VANDENDIJCK et al., 2024), onde a superfície é modelada em função da idade do respondente e da *diferença* de idade ($d_{ij} = x_i - y_j$) para capturar padrões geracionais (como a interação entre pais e filhos), refletindo o padrão de "assortatividade etária" e permitindo filtrar o ruído amostral

sem perder os principais contornos de interação (KASSTEELE; EIJKEREN; WALLINGA, 2017).

A seleção do modelo final não se baseou apenas em métricas de ajuste interno (Deviance Explained, AIC, k-index), mas foi decidida por um critério de validação externa: conduzimos uma validação cruzada 5-fold sobre os pares (i, j), calculando o CV-RMSE (Root Mean Squared Error da validação cruzada) das taxas preditas versus observadas. Este critério foi decisivo, pois penaliza naturalmente a complexidade excessiva e mede a capacidade de generalização do modelo. Complementarmente, a avaliação qualitativa foi crucial para o diagnóstico final, utilizando a inspeção de *heatmaps* de resíduos (para garantir que não havia viés estrutural sistemático), gráficos de dispersão (Resíduos vs. Valores Ajustados, para validar a estabilidade da variância e a correção da heterocedasticidade), e *Q-Q Plots* (para verificar a aderência dos resíduos padronizados à distribuição teórica).

2 Metodologia e Análise Exploratória de Dados

2.1 Delineamento e Caracterização da Amostra

O presente estudo adota uma abordagem quantitativa baseada na modelagem estatística de matrizes de contato por idade para o Complexo de Manguinhos (RJ) . Utilizam-se dados observacionais provenientes do inquérito COMVIDA-Fiocruz, composto por 4.033 respondentes.

A amostra apresentou mediana de idade de 39,8 anos, com 21,7% dos participantes tendo 60 anos ou mais e uma predominância do sexo feminino (60,7%). O perfil dos moradores reforça a vulnerabilidade social do território: 40,7% da população autodeclarou-se parda e 22,7% negra; 44% possuem menos de 10 anos de escolaridade e 40,4% das famílias sobrevivem com renda mensal de até um salário mínimo. Adicionalmente, as condições estruturais incluem alta densidade domiciliar (69,8% dos domicílios com três ou mais moradores) e alta dependência de transporte coletivo (43,6% dos participantes).

2.1.1 Viés de Seleção e Ponderação (w_i)

Ao compararmos a estrutura etária da amostra com os dados censitários (IBGE 2010), nota-se um viés de seleção etária evidente, com sub-representação de jovens e super-representação de idosos (ver Figura 20 no Apêndice). Para mitigar esse viés e garantir representatividade, definiu-se o peso amostral w_i:

$$w_i = \frac{P_i}{r_i} \tag{2.1}$$

onde P_i é o tamanho da população censitária na faixa i e r_i é o número de respondentes naquele grupo. Este peso é utilizado nas etapas subsequentes para corrigir a sub ou super-representação etária da amostra.

2.2 Definição de Variáveis e Tratamento do Sub-relato

O banco de dados inclui a idade dos respondentes, idade dos contatos reportados, número de moradores e número de interações diárias. Um "contato" foi definido para capturar interações epidemiologicamente relevantes, como contato físico pele a pele ou conversa bidirecional próxima.

A análise bruta revelou, contudo, um desafio metodológico: 64,38% dos entrevistados relataram *zero* contatos no dia anterior, um número incompatível com a alta densidade

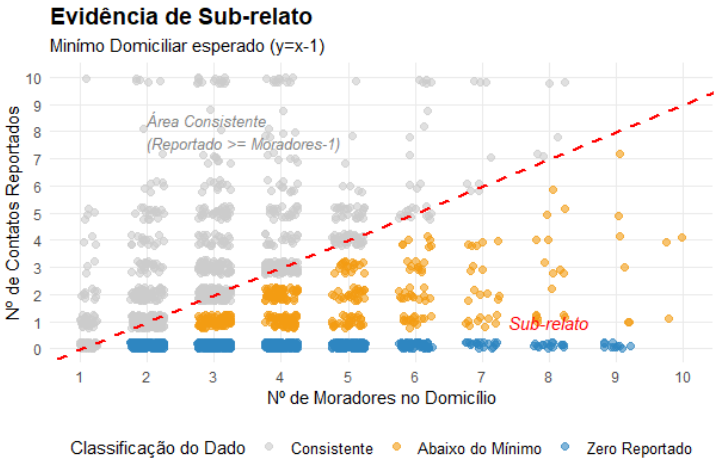
domiciliar observada. A análise da distribuição bruta dos contatos (ver Figura 21 no Apêndice) revela um cenário de dados inflacionados por zeros e com alta variabilidade. A média de contatos é baixa $(1,1)$, porém a variância é significativamente maior $(4,28)$, resultando em uma razão de dispersão (*dispersion ratio*) de $3,9$. Esses indicadores confirmam a sobredispersão dos dados, sugerindo que modelos baseados em Poisson simples seriam insuficientes e justificando a adoção posterior de abordagens como a Binomial Negativa.

2.2.1 Correção Metodológica pelo Mínimo Domiciliar

Os participantes responderam à pergunta 58: *"Você teve contato dentro ou fora de sua casa (conversou mais de três frases, tocou ou foi tocado) com alguém ontem?"*. Embora o enunciado da pergunta inclua explicitamente o ambiente doméstico ("dentro de sua casa"), a discrepância entre os relatos de zero contatos e a realidade de domicílios superlotados sugere um viés sistemático. É provável que os respondentes tenham interpretado "contato" apenas como interações sociais externas ou extraordinárias, desconsiderando a convivência rotineira com coabitantes.

Entretanto, em um contexto epidemiológico, especialmente sob isolamento ou distanciamento social, a exposição intradomiciliar é contínua e inevitável. Logicamente, espera-se que um indivíduo interaja, no mínimo, com os demais moradores de sua residência. Ignorar esse fato subestimaria drasticamente o risco de transmissão secundária nos lares.

Figura 1 – Evidência de Sub-relato



Fonte: Elaboração da autora.

Para corrigir essa distorção e impor a consistência lógica da transmissão intradomiciliar, adotou-se a seguinte regra de correção para o cálculo do número de contatos (contact_count):

$$\text{contact\_count} = \max\{\text{contatos reportados}, \text{moradores no lar} - 1\} \tag{2.2}$$

Tal critério assegura que sejam contabilizados todos os contatos domiciliares. A necessidade e eficácia dessa intervenção são demonstradas na Figura 22 (Apêndice), que compara as médias de contato por idade antes e depois do ajuste. Observa-se que os dados brutos (pontos vermelhos) apresentavam uma distribuição achatada e biologicamente implausível, com crianças reportando médias próximas a 1 contato diário. Após a correção (pontos verdes), recupera-se o padrão de mistura esperado: as faixas etárias mais jovens passam a exibir médias superiores a 3, refletindo a intensa interação intradomiciliar característica desses grupos, enquanto a curva global assume um comportamento decrescente com a idade, condizente com a literatura de contatos sociais.

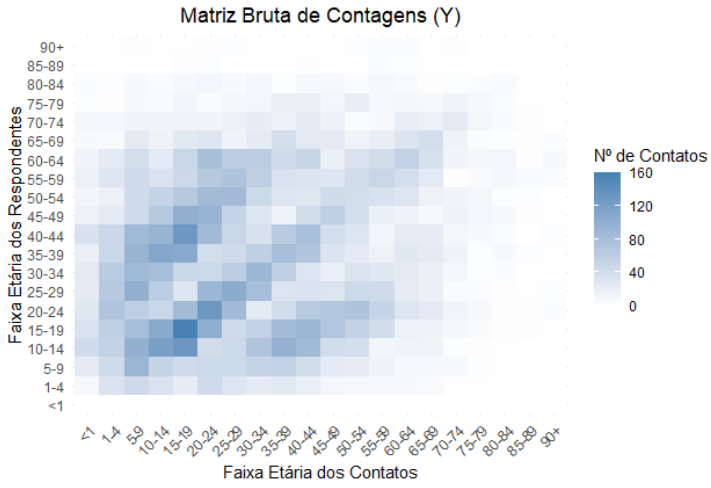
2.3 Formalização Matemática das Matrizes de Contato

Para a construção das matrizes, as idades foram estratificadas em $m = 20$ faixas etárias padronizadas ($< 1, 1-4, \ldots, 90+$), onde cada faixa i é representada pelo ponto médio x_i. O domínio das matrizes é definido por $i, j \in \{1, 2, \ldots, m\}$.

2.3.1 Matriz de Contagens (Y)

Define-se a matriz de contagens brutas $Y = [Y_{ij}]_{m \times m}$, onde $Y_{ij} \in \mathbb{N}_0$ representa o número total de contatos observados entre respondentes da faixa etária i e contatos na faixa j.

Figura 2 – Matriz Bruta de Contagens (Y)



Fonte: Elaboração da autora.

2.3.2 Matriz de Exposições (E) e Taxas (Γ)

A matriz de exposições é definida por $E = [E_{ij}]_{m\times m}$, onde $E_{ij} = r_i$ (número de respondentes na faixa i). A partir destas, obtém-se a **Matriz de Taxas Brutas (Γ)**:

$$\Gamma = [\Gamma_{ij}]_{m\times m}, \quad \text{onde} \quad \Gamma_{ij} = \frac{Y_{ij}}{E_{ij}} \tag{2.3}$$

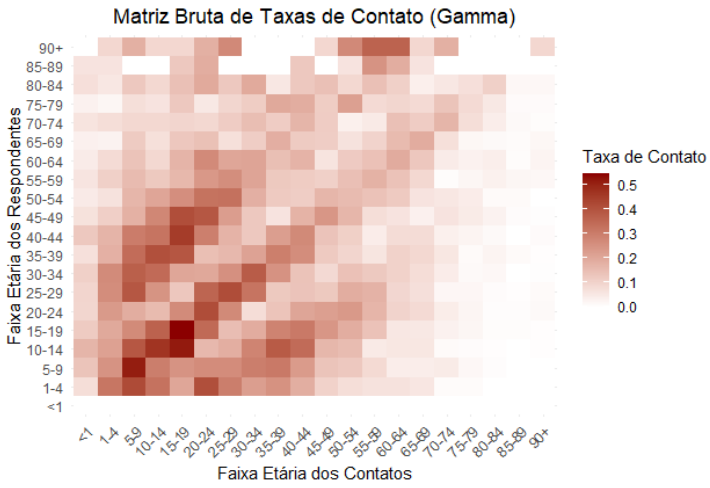
Interpretação Probabilística

A taxa Γ_{ij} é interpretada como:

$$\Gamma_{ij} \approx E[C \mid A = i, A' = j] \tag{2.4}$$

onde C é o número de contatos, A é a idade do respondente e A' é a idade do contato.

Figura 3 – Matriz Bruta de Taxas de Contato (Γ)



Fonte: Elaboração da autora.

A visualização da matriz bruta Γ (Figura 3) expõe características essenciais da dinâmica social. Nota-se uma **forte assortatividade etária** (homofilia), evidenciada pela diagonal principal marcada, indicando que indivíduos interagem predominantemente com pares da mesma idade. Padrões fora da diagonal (diagonais secundárias) também são visíveis, capturando misturas intergeracionais típicas, como as interações entre pais e filhos ou avós e netos, cruciais em contextos de domicílios multigeracionais.

Contudo, a matriz bruta apresenta limitações críticas para uso direto em modelagem epidemiológica. Observa-se **ruído estocástico** e **esparsidade** (células vazias ou com taxas abruptas), que são artefatos da variabilidade amostral e não ausência real de interação. Conforme discutido por Vandendijck et al. (2024), o comportamento de contato social muda de forma contínua e gradual com o envelhecimento, e não de maneira discreta ou abrupta entre faixas etárias. Portanto, a aplicação de métodos de suavização (como LOESS e GAM) torna-se indispensável para: (1) filtrar o ruído amostral, recuperando a superfície

contínua subjacente que descreve a verdadeira dinâmica populacional; e (2) preencher lacunas de informação, garantindo estimativas robustas de parâmetros epidemiológicos (como R_0) que não sejam enviesadas por flutuações aleatórias de uma amostra finita.

2.4 Suavização via LOESS Bidimensional

A técnica LOESS (*Locally Estimated Scatterplot Smoothing*) foi empregada como uma primeira abordagem não-paramétrica para estimar a superfície contínua subjacente $f : \mathbb{R}^2 \to \mathbb{R}_+$ tal que $f(x_i, y_j) \approx \Gamma_{ij}$. Este método ajusta modelos de regressão local em torno de cada ponto da grade, permitindo capturar a estrutura de dependência local dos dados sem impor uma forma funcional global rígida.

2.4.1 Formulação do Problema de Minimização

Para cada ponto alvo (x_i, y_j) na grade de idades, a estimativa suavizada $\hat{f}(x_i, y_j)$ é obtida pela solução de um problema de mínimos quadrados ponderados locais. O estimador busca os coeficientes β que minimizam a seguinte função objetivo em uma vizinhança local:

$$\hat{f}(x_i, y_j) = \arg\min_{\beta_0,\beta_1,\beta_2} \sum_{u=1}^{m}\sum_{v=1}^{m} w_u K\left(\frac{d_{(i,j),(u,v)}}{h}\right) [\Gamma_{uv} - \beta_0 - \beta_1(x_u - x_i) - \beta_2(y_v - y_j)]^2 \tag{2.5}$$

Nesta formulação:

- **Kernel de Suavização $K(\cdot)$:** Utilizou-se uma função de peso tricúbica, definida como $K(d) = (1 - |d|^3)^3$ para $|d| < 1$ e 0 caso contrário. Essa escolha é fundamental porque ela decai suavemente para zero nas bordas da vizinhança, o que garante a continuidade da superfície suavizada e melhora a aproximação estatística da variância do erro.

- **Parâmetro de Banda h (Span):** A escolha do parâmetro *span* buscou equilibrar o compromisso fundamental entre viés e variância, utilizando como critério objetivo os Graus de Liberdade Efetivos (EDF), definidos pelo traço da matriz de projeção (*hat matrix*). Através de uma varredura computacional (ver Figura 23), o parâmetro foi calibrado em 0,08 para atingir aproximadamente 40 graus de liberdade ($2m$). Este alvo é justificado pela dimensão 20×20 da matriz de contatos, permitindo modelar os efeitos principais de idade e coorte com flexibilidade suficiente para evitar o *overfitting* do ruído estocástico das células individuais.

- **Ponderação Demográfica Normalizada w_u:** Os pesos demográficos iniciais, dados por P_u/r_u, foram normalizados pela sua média global antes do ajuste. Esse

procedimento garante a estabilidade numérica do algoritmo ao manter a soma dos pesos proporcional ao tamanho da amostra, evitando distorções na escala da função objetivo e na estimativa da variância residual.

2.4.2 Definição de Distância e Estruturas de Modelo

A distância $d_{(i,j),(u,v)}$ define a geometria da vizinhança e foi modelada de duas formas distintas para capturar diferentes dinâmicas de interação:

1. **Modelo Idade-Idade (M1):** Utiliza a distância euclidiana direta no plano cartesiano das idades:
$$d_{(i,j),(u,v)} = \sqrt{(x_u - x_i)^2 + (y_v - y_j)^2} \tag{2.6}$$

2. **Modelo de Coorte (M2):** Para enfatizar padrões geracionais (como a interação pais-filhos), introduz-se a variável de diferença etária $d_{ij} = x_i - y_j$. A distância é então calculada no plano transformado (x, d):
$$d_{(i,j),(u,v)} = \sqrt{(x_u - x_i)^2 + (d_{uv} - d_{ij})^2} \tag{2.7}$$
Esta reformulação alinha a suavização ao longo das diagonais da matriz, capturando a estrutura de coorte inerente ao envelhecimento populacional.

Por fim, dado que a regressão local linear pode predizer valores negativos, aplicou-se um truncamento a posteriori $\hat{f} \leftarrow \max\{\hat{f}, \epsilon\}$ para garantir a coerência biológica das taxas, onde ϵ foi definido como metade do menor valor positivo predito pelo modelo ($\epsilon = \hat{f}_{min>0}/2$).

2.5 Modelos Aditivos Generalizados (GAM)

Embora o LOESS forneça uma aproximação visual útil, ele trata as taxas como variáveis contínuas com erros simétricos, o que é inadequado para dados de contatos sociais que são, por natureza, contagens discretas com muitos zeros e alta variabilidade. Neste contexto, os Modelos Aditivos Generalizados (GAM) oferecem uma abordagem superior e mais robusta. Eles modelam explicitamente a natureza estocástica da contagem Y_{ij} (números inteiros não negativos), acomodam naturalmente o excesso de zeros e lidam corretamente com a relação média-variância (heterocedasticidade) típica de dados epidemiológicos, onde a variância cresce com a média.

2.5.1 Estrutura do Modelo e Preditores Lineares

Assumimos que a contagem Y_{ij} segue uma distribuição da família exponencial com média condicional $\mu_{ij} = E(Y_{ij}|x_i, y_j)$. O modelo relaciona essa média aos preditores através

de uma função de ligação logarítmica, garantindo que as taxas preditas sejam sempre positivas:

$$\log(\mu_{ij}) = \log(E_{ij}) + s(x_i, y_j) \tag{2.8}$$

- **Offset:** O termo $\log(E_{ij})$ é incluído como um *offset* fixo. Isso normaliza a contagem pela exposição (número de respondentes r_i), permitindo que o modelo estime a taxa de contato intrínseca enquanto respeita a distribuição discreta dos dados de contagem (LITVINOVA et al., 2025).

- **Superfície Suave** $s(\cdot, \cdot)$**:** A interação não linear entre as idades foi modelada utilizando *tensor product splines* (splines de produto tensorial). Esta técnica é crucial porque, diferente de abordagens que impõem o mesmo nível de suavidade em todas as direções (isotropia), ela acomoda a variação desigual da estrutura da matriz de contatos. A superfície de contato apresenta picos agudos e estreitos na diagonal principal (assortatividade) que coexistem com platôs mais suaves nas regiões de interação intergeracional, exigindo uma flexibilidade diferente para o eixo do respondente e o eixo do contato (LITVINOVA et al., 2025).

Essa construção bidimensional é feita através do produto de bases marginais unidimensionais (neste estudo, *thin-plate regression splines* com penalização *shrinkage*), conforme a equação:

$$s(x_i, y_j) = \sum_{l=1}^{K_x} \sum_{k=1}^{K_y} \beta_{lk} b_l(x_i) c_k(y_j) \tag{2.9}$$

Nesta formulação, $b_l(x_i)$ e $c_k(y_j)$ são as funções de base marginais para respondente e contato, respectivamente. Os termos β_{lk} representam os coeficientes a serem estimados pelo modelo e as constantes K_x, K_y definem a dimensão da base, controlando o teto de flexibilidade permitido para a superfície em cada eixo. A eficácia desta estrutura foi avaliada tanto na configuração padrão de idades (x_i, y_j) quanto na configuração rotacionada de coorte (x_i, d_{ij}), visando otimizar a detecção dos efeitos diagonais (VANDENDIJCK et al., 2024).

2.5.2 Estimação por Verossimilhança Penalizada

A estimação dos coeficientes β não busca apenas o ajuste aos dados, mas penaliza a complexidade excessiva (rugosidade) da superfície. O critério de maximização é a log-verossimilhança penalizada:

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_r \lambda_r \beta^\top S_r \beta \tag{2.10}$$

onde $l(\beta)$ é a log-verossimilhança da família de distribuição adotada, S_r são matrizes de penalização associadas à rugosidade da superfície, e λ_r são os parâmetros de suavização.

Estes parâmetros foram estimados via REML (*Restricted Maximum Likelihood*), garantindo um compromisso ótimo e automático entre o viés do ajuste e a variância da estimativa (VANDENDIJCK et al., 2024).

2.5.3 Famílias de Distribuição

Foram testadas diferentes famílias de distribuição para modelar a variável resposta Y_{ij}:

- **Poisson:** Assume que a variância é igual à média ($Var(Y_{ij}) = \mu_{ij}$). Esta família foi rejeitada devido à forte evidência de sobredispersão observada na análise exploratória ($Var(Y) > E(Y)$), o que indica uma variabilidade nos dados brutos muito superior àquela que o modelo Poisson consegue capturar.
- **Quasi-Poisson:** Modela a variância como $Var(Y_{ij}) = \phi\mu_{ij}$, onde $\phi > 1$ é um parâmetro de dispersão constante estimado a posteriori. Embora corrija a inferência (erros-padrão), não altera a estrutura probabilística da média.
- **Binomial Negativa (NB):** Adota uma relação média-variância quadrática $Var(Y_{ij}) = \mu_{ij} + \mu_{ij}^2/\theta$, onde θ é o parâmetro de dispersão adicional. Esta formulação é teoricamente mais adequada para capturar a heterogeneidade excessiva e a agregação típica de dados de contatos sociais

2.6 Critérios de Avaliação e Seleção de Modelos

A avaliação da qualidade das matrizes de contato geradas requer uma abordagem híbrida, combinando métricas escalares de ajuste com diagnósticos visuais dos resíduos. Como o método LOESS baseia-se em regressão local e os GAMs em verossimilhança penalizada, definem-se abaixo os critérios quantitativos e qualitativos adotados para a seleção do modelo final.

2.6.1 Métricas para Avaliação do LOESS (Modelos M1 e M2)

Como o LOESS não possui uma função de verossimilhança explícita, a comparação entre o Modelo Padrão (M1) e o Modelo de Coorte (M2) baseia-se na distância euclidiana entre os valores ajustados e os observados.

1. **Raiz do Erro Quadrático Médio (RMSE - In-sample)**

   O RMSE quantifica a magnitude média do erro de ajuste nos mesmos dados utilizados para treinar o modelo, sendo calculado por:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i,j}(\Gamma_{ij} - \hat{\Gamma}_{ij})^2} \tag{2.11}$$

Esta métrica avalia a distância média absoluta entre as taxas de contato brutas (Γ_{ij}) e a superfície suavizada ($\hat{\Gamma}_{ij}$), sendo utilizada para verificar qual estrutura geométrica adere melhor aos dados observados: a cartesiana direta (M1) ou a transformada por coorte (M2). Espera-se que modelos com maior flexibilidade apresentem menor RMSE, embora um valor baixo isoladamente não garanta capacidade preditiva.

2. **RMSE de Validação Cruzada (CV-RMSE)**

   Para avaliar a capacidade de generalização e evitar o *overfitting*, adota-se um esquema de validação cruzada *k-fold* (com $k = 5$), calculando o erro apenas nos dados de teste:

$$CV\text{-}RMSE = \sqrt{\frac{1}{N_{teste}} \sum_{k \in Teste} (\Gamma_k - \hat{\Gamma}_k)^2} \tag{2.12}$$

   Ao calcular o erro sobre taxas de contato desconhecidas pelo modelo, esta métrica atua como o critério de desempate entre M1 e M2. Ela revelará se a imposição da estrutura de coorte no M2 captura de fato um padrão populacional robusto ou se apenas memoriza ruídos da amostra, sendo escolhido o modelo que minimizar o CV-RMSE.

2.6.2 Métricas para Avaliação dos GAM (Modelos M3 a M6)

Para os Modelos Aditivos Generalizados, a seleção envolve a escolha da família de distribuição (Poisson, Quasi-Poisson ou Binomial Negativa) e a verificação da dimensão da base (k).

1. **Deviância Explicada (Deviance Explained)**

   Análoga ao R^2 da regressão linear, esta métrica mensura a proporção da variabilidade dos dados que é capturada pelo modelo em relação a um modelo nulo:

$$DevExp = \frac{D_{nulo} - D_{residual}}{D_{nulo}} \times 100\% \tag{2.13}$$

   Ela é essencial para monitorar o ganho de ajuste conforme se aumenta a complexidade da modelagem, permitindo comparar o desempenho do modelo Poisson (M3) contra as variantes mais robustas que lidam com a dispersão dos dados, especificamente a Binomial Negativa (M4 e M6) e a Quasi-Poisson (M5).

2. **Critério de Informação de Akaike (AIC)**

   O AIC é uma medida comparativa que balança a bondade de ajuste com a complexidade do modelo, penalizando o número de parâmetros estimados:

$$AIC = 2k - 2\ln(\hat{L}) \tag{2.14}$$

Este critério é fundamental para justificar a escolha da família de distribuição, determinando se a introdução de parâmetros extras para modelar a sobredispersão — presentes na Binomial Negativa e na estrutura de variância da Quasi-Poisson — é estatisticamente justificável e compensa o custo de complexidade frente ao modelo Poisson simples.

3. **Índice de Checagem de Base (k-index)**

   Este diagnóstico, obtido via função `gam.check`, verifica se a dimensão da base das splines (k) é suficiente para capturar a rugosidade da superfície:

$$\text{k-index} = \frac{\text{Variância Residual Estimada}}{\text{Variância Residual Teórica}} \tag{2.15}$$

   O índice serve como critério técnico para validar a escolha do número de nós das splines, indicando se restam padrões sistemáticos nos resíduos que o modelo falhou em capturar. Ele será usado para decidir entre as bases 15×15 e 20×20 nos modelos M4 e M6, garantindo que a suavização não esteja excessivamente rígida.

4. **Comparação via CV-RMSE para GAM**

   Assim como no LOESS, o erro de validação cruzada é calculado para as predições dos modelos probabilísticos, servindo como o "juiz final" para detectar *overfitting*. Esta métrica será usada especificamente para comparar o modelo padrão (M4) com o modelo de coorte (M6) sob diferentes dimensões de base ($k = 15$ vs $k = 20$). Se o aumento de k melhorar o AIC mas piorar o CV-RMSE, opta-se pela opção mais parcimoniosa.

2.6.3 Diagnóstico Visual e Qualitativo (Heatmaps e Resíduos)

Enquanto as métricas quantitativas fornecem um resumo escalar do ajuste, a análise qualitativa visual permite identificar onde e como o modelo falha. Foram construídos três tipos de gráficos de diagnóstico para validar as premissas estatísticas.

1. **Heatmap de Resíduos (Diagnóstico Estrutural)**

   Este gráfico projeta a matriz de erros $\epsilon_{ij} = \Gamma_{ij} - \hat{\Gamma}_{ij}$ no plano *Idade Respondente* $\times$ *Idade Contato*, onde regiões em vermelho escuro indicam subestimação e regiões em azul indicam superestimação. Um bom modelo deve apresentar um padrão visual de "ruído branco" (cores claras e dispersas aleatoriamente), validando um alto *Deviance Explained*, enquanto a presença de blocos contíguos de cor forte indicaria viés sistemático não capturado pelas splines.

2. **Gráfico de Dispersão: Resíduos vs. Valores Ajustados (Diagnóstico de Variância)**

Plota-se os valores ajustados pelo modelo ($\hat{y}$) contra os resíduos (de Pearson para GAM e ordinários para LOESS). Este gráfico é decisivo para a escolha da família de distribuição: se a nuvem formar um "cone" abrindo para a direita, evidencia-se sobredispersão, invalidando o modelo Poisson; já uma dispersão retangular e uniforme valida que o parâmetro de dispersão (θ) da Binomial Negativa ou da Quasi-Poisson controlou adequadamente a heterocedasticidade.

3. **Q-Q Plot (Diagnóstico Distribucional)**

   Compara os quantis dos resíduos padronizados (Resíduos de *Deviance* para GAM) contra os quantis teóricos de uma distribuição Normal. A aderência dos pontos à linha diagonal vermelha ($y = x$) valida a função de ligação e a distribuição de erro escolhida. Desvios severos nas caudas sugerem inadequação da família de distribuição ou má gestão de *outliers*, o que comprometeria a confiabilidade dos intervalos de confiança e dos testes de hipótese.

3 Resultados

3.1 Avaliação da Suavização Não-Paramétrica (LOESS)

A análise visual das matrizes suavizadas revela diferenças estruturais marcantes impostas pela geometria de cada modelo. A Figura 3.1 apresenta as taxas de contato suavizadas $\hat{\Gamma}_{ij}$ resultantes lado a lado.

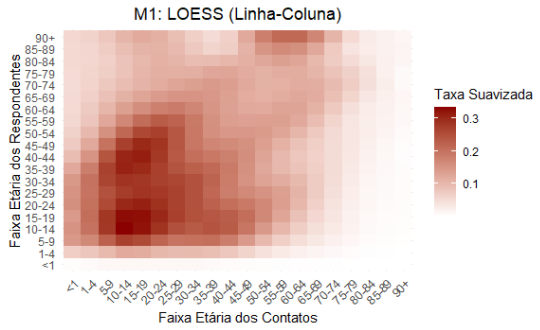


Figura 4 – Matriz suavizada M1.

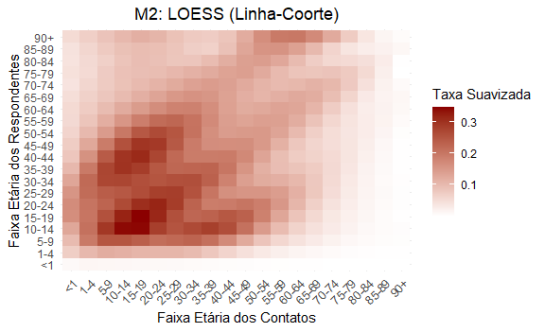


Figura 5 – Matriz suavizada M2.

Fonte: Elaboração da autora.

O modelo M1 tende a gerar padrões retangulares ("efeito xadrez"), suavizando as taxas predominantemente nas direções horizontal e vertical. Embora capture a alta densidade de contatos em crianças e jovens, ele falha em delinear com clareza as interações que ocorrem ao longo das diagonais da matriz.

Em contraste, o modelo M2, ao incorporar a diferença de idade d_{ij} como preditor, produz uma superfície com forte orientação diagonal. Isso resulta em uma representação biologicamente mais plausível da assortatividade (a faixa diagonal principal intensa, indicando interação entre pares da mesma idade) e das misturas intergeracionais (diagonais secundárias, refletindo interações pais-filhos). A suavização por coorte "respeita" a dinâmica temporal do envelhecimento, preservando estruturas que o M1 dilui.

3.1.1 Desempenho Métrico (In-Sample vs. Out-of-Sample)

Conforme observado na Tabela 1, o Modelo de Coorte (M2) apresentou um desempenho métrico superior.

Tabela 1 – Métricas de desempenho para os modelos LOESS M1 e M2.

Modelo	**RMSE (In-Sample)**	**CV-RMSE**
M1 (Linha-Coluna)	0.0626	0.0669
M2 (Linha-Coorte)	0.0598	0.0646

Fonte: Elaboração da autora.

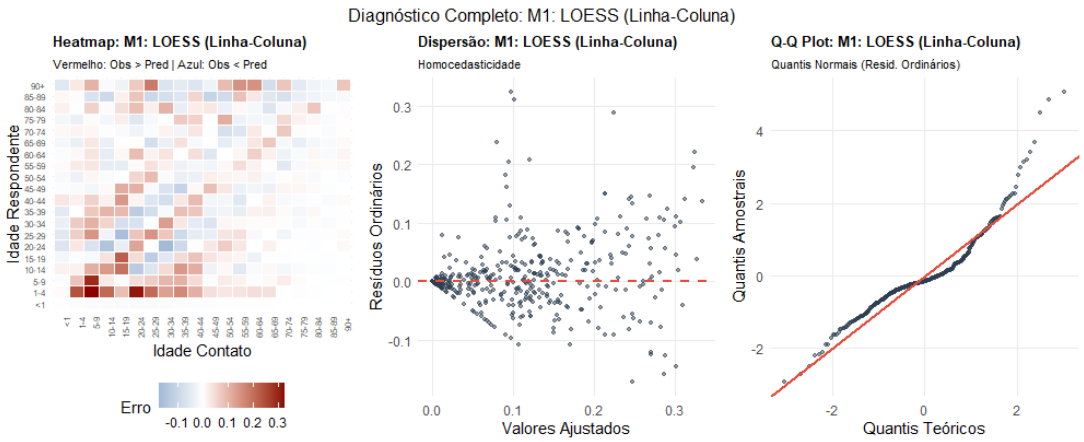
O RMSE In-Sample de 0,0598 indica que o M2 se ajustou melhor aos dados observados do que o M1 (0,0626). Mais importante, o CV-RMSE (0,0646) do M2 também foi inferior ao do M1 (0,0669), sugerindo que a imposição da estrutura de coorte não apenas melhorou o ajuste aos dados de treino, mas também a capacidade de prever taxas de contato em subconjuntos de dados não observados.

Apesar da vantagem numérica e visual do M2, a diferença de magnitude nos erros é pequena, o que indica a necessidade de verificar se os resíduos do modelo se comportam de maneira adequada antes de validá-lo.

3.1.2 Diagnóstico de Resíduos e Limitações do Método

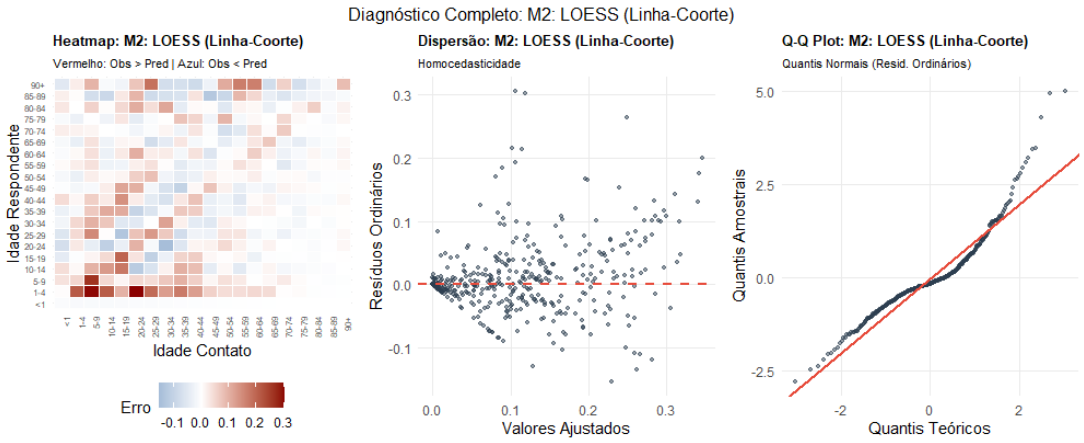
A análise crítica dos diagnósticos de resíduos revela a inadequação fundamental do LOESS para este tipo de dado. As Figuras abaixo apresentam o painel diagnóstico completo para M1 e M2.

Figura 6 – Diagnóstico completo do Modelo M1.



Fonte: Elaboração da autora.

Figura 7 – Diagnóstico completo do Modelo M2.



Fonte: Elaboração da autora.

A análise dos gráficos de Resíduos Ordinários vs. Valores Ajustados (painel central nas Figuras 6 e 7) expõe uma violação crítica da hipótese de homocedasticidade. Diferente de um "bom ajuste", onde os resíduos formariam uma faixa retangular uniforme em torno do zero, observa-se aqui um padrão de abertura em leque (cone abrindo para a direita): para taxas de contato baixas, os erros são pequenos e concentrados; porém, à medida que a taxa de contato estimada aumenta, a dispersão dos erros cresce drasticamente.

Este comportamento confirma que a variância não é constante, mas sim proporcional à média, o que contradiz a premissa fundamental do LOESS de que os erros são independentes da magnitude do contato. Adicionalmente, os Q-Q Plots (painel à direita) mostram desvios significativos nas caudas, confirmando que a distribuição dos erros não segue uma Normalidade. Isso justifica a necessidade de modelos probabilísticos (GAM) que lidem nativamente com essa estrutura de variância.

3.2 Avaliação dos Modelos Probabilísticos: Poisson vs. Quasi-Poisson

A transição para os Modelos Aditivos Generalizados (GAM) teve como objetivo corrigir as limitações estatísticas observadas na suavização LOESS, especificamente a incapacidade de lidar com a natureza de contagem discreta e a heterocedasticidade dos dados. Inicialmente, avaliamos o modelo Poisson (M3) e sua variante Quasi-Poisson (M4) para verificar o ajuste à dispersão dos dados.

As Figuras 3.2 apresentam o impacto visual da mudança de família de distribuição na superfície de contato estimada.

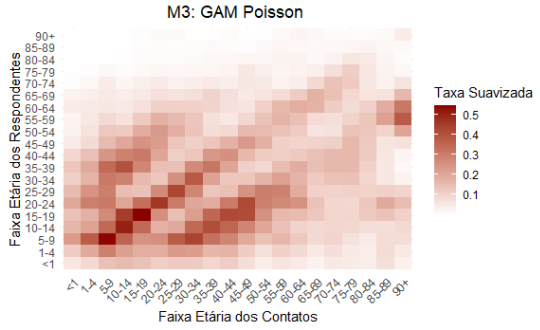


Figura 8 – Matriz suavizada M3.

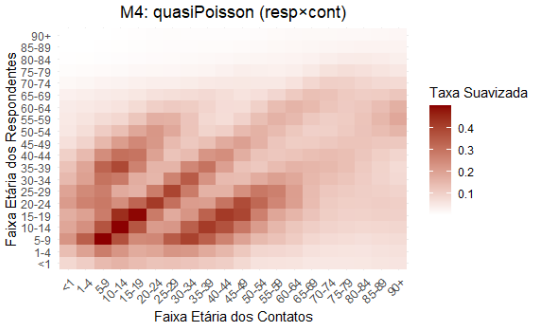


Figura 9 – Matriz suavizada M4.

Fonte: Elaboração da autora.

3.2.1 Diagnóstico Comparativo de Ajuste e Dispersão

A Tabela 2 resume as métricas de desempenho obtidas para os dois modelos, mantendo a dimensão da base de suavização fixa em $k = 15$.

Tabela 2 – Comparativo de desempenho: Poisson (M3) vs. Quasi-Poisson (M4). A estimativa de escala elevada no M4 evidencia a severa sobredispersão ignorada pelo M3.

Modelo	**Família**	**Escala ($\hat{\phi}$)**	**EDF Total**	**Dev. Exp. (%)**	**k-index (p-valor)**
M3	Poisson	1 (fixo)	190,3	94,6%	1,01 (0,74)
M4	Quasi-Poisson	16,06	93,4	93,0%	1,02 (0,80)

Fonte: Elaboração da autora.

Nota: EDF = Graus de Liberdade Efetivos. O k-index > 1 *indica que a dimensão da base* $k = 15$ *é suficiente.*

Análise de Sobredispersão

O modelo Poisson (M3) assume rigidamente que a variância é igual à média ($Var(Y) = \mu$), fixando o parâmetro de escala em 1. No entanto, ao relaxar essa restrição com o modelo Quasi-Poisson (M4), a estimativa do parâmetro de dispersão resultou em $\hat{\phi} \approx 16,06$. Isso indica que a variabilidade real dos contatos é mais de 16 vezes superior ao previsto pela distribuição de Poisson, confirmando um quadro de **sobredispersão severa**.

A consequência dessa inadequação do modelo Poisson é visível na coluna **EDF** (Graus de Liberdade Efetivos). O M3 utilizou 190,3 graus de liberdade (próximo do máximo possível para $k = 15$), o que sugere um modelo excessivamente flexível que está tentando ajustar o ruído estocástico dos dados como se fosse sinal estrutural.

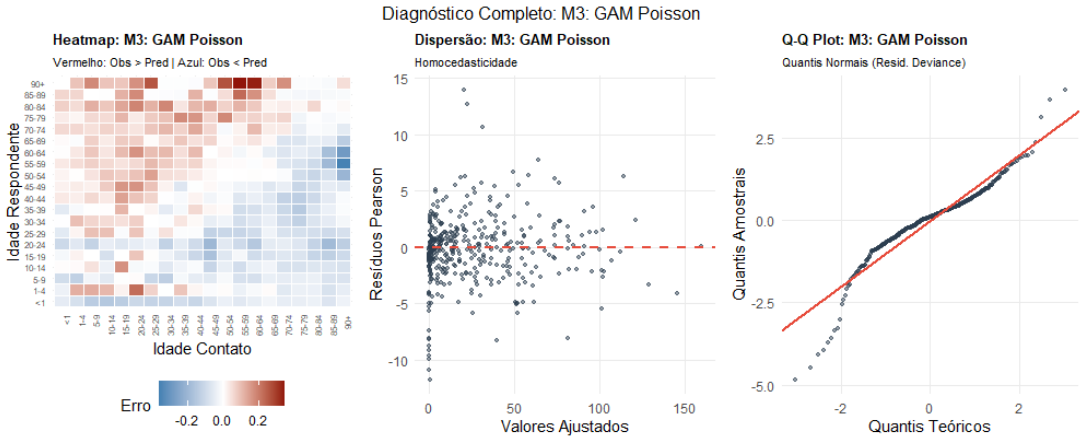
Por outro lado, o modelo Quasi-Poisson, ao incorporar a variância extra ($\hat{\phi} \approx 16$), penalizou drasticamente a complexidade da superfície, reduzindo o EDF para 93,4. Embora

estatisticamente mais correto, essa redução resultou em um **subajuste (oversmoothing)**, onde a superfície suavizada (Figura 9) perdeu detalhes importantes da estrutura de contato, aparecendo mais "borrada" em comparação ao M3.

3.2.2 Diagnóstico Visual de Resíduos

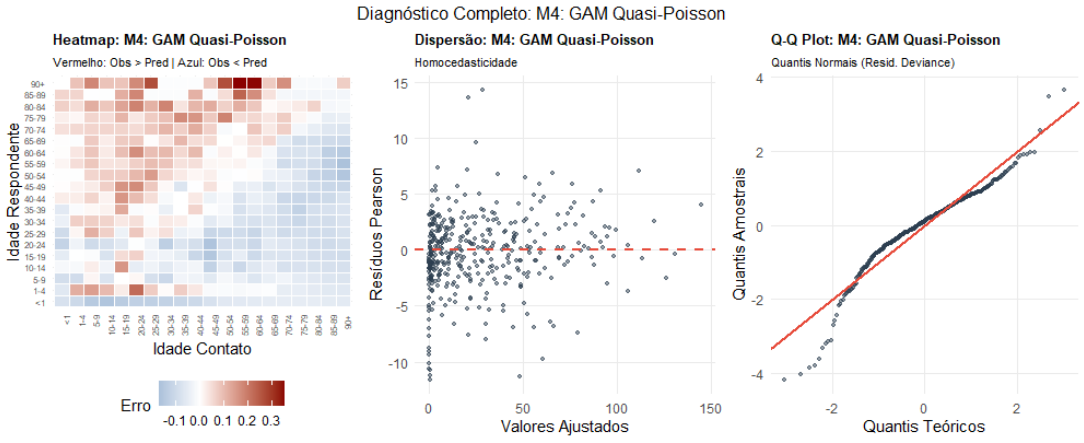
A inspeção visual dos resíduos corrobora com as métricas da tabela e expõe por que a família Poisson é inadequada para dados de contato social.

Figura 10 – Diagnóstico do Modelo GAM Poisson (M3).



Fonte: Elaboração da autora.

Figura 11 – Diagnóstico do Modelo GAM Quasi-Poisson (M4).



Fonte: Elaboração da autora.

Nas Figuras 10 e 11, o gráfico de **Resíduos vs. Valores Ajustados** (painel central) apresenta, mais uma vez, um padrão de dispersão crescente, indicando que a variabilidade não é constante ao longo da amostra. Isso demonstra que, embora o Quasi-Poisson ajuste

os erros-padrão, ele não altera a relação funcional média-variância subjacente, falhando em estabilizar a variância para valores preditos mais altos.

Além disso, ambos os modelos passaram no teste de verificação de base (`gam.check`), apresentando *k-index* > 1 e p-valores altos (0,74 e 0,80), indicando que a limitação não está no número de funções de base ($k = 15$), mas sim na escolha incorreta da família de distribuição.

Portanto, a rejeição da família Poisson (por sobredispersão e superajuste) e as limitações da Quasi-Poisson (penalização excessiva da suavidade) apontam para a necessidade de uma distribuição que modele explicitamente a heterogeneidade dos dados. A Binomial Negativa, que possui um parâmetro de forma para acomodar a variância quadrática ($Var(Y) = \mu + \mu^2/\theta$), apresenta-se como a próxima candidata natural para equilibrar ajuste e complexidade.

3.3 Avaliação dos Modelos Probabilísticos: Binomial Negativa

Após rejeitar as famílias Poisson e Quasi-Poisson devido à inadequação no tratamento da dispersão, a análise focou na família Binomial Negativa (NB). Esta etapa avaliou simultaneamente duas dimensões de complexidade: a estrutura geométrica do preditor (Padrão vs. Coorte) e a flexibilidade da base de suavização ($k = 15$ vs. $k = 20$).

As Figuras 3.3 e 3.3 ilustram o impacto do aumento da complexidade da base nas superfícies de contato estimadas pelos modelos M5 (Padrão) e M6 (Coorte), respectivamente.

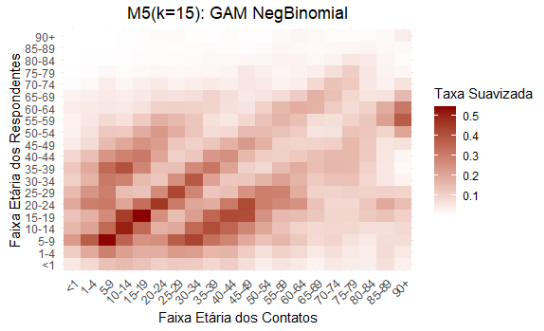


Figura 12 – Matriz suavizada M5 ($k = 15$).

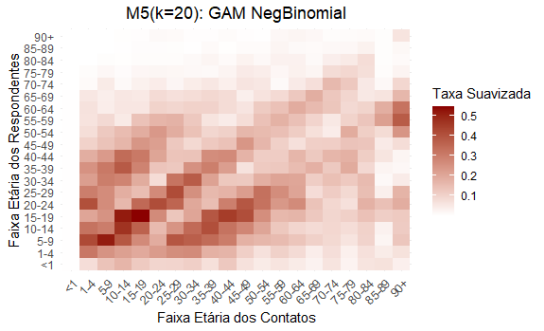


Figura 13 – Matriz suavizada M5 ($k = 20$).

Fonte: Elaboração da autora.

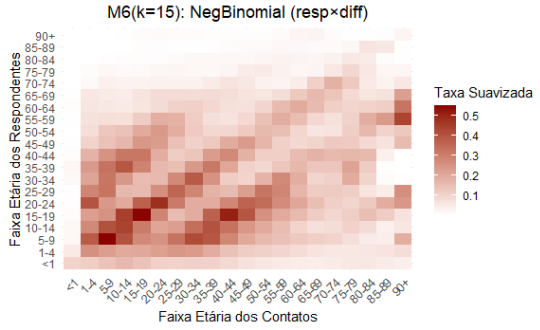


Figura 14 – Matriz suavizada M6 ($k = 15$).

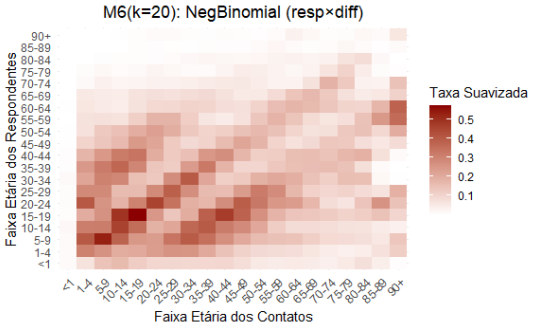


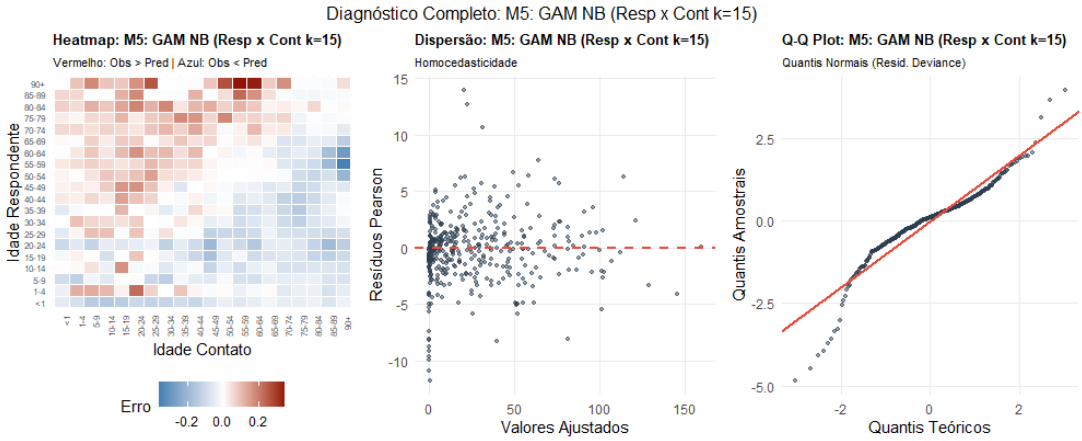
Figura 15 – Matriz suavizada M6 ($k = 20$).

Fonte: Elaboração da autora.

3.3.1 Diagnóstico de Ajuste e Estabilidade da Variância

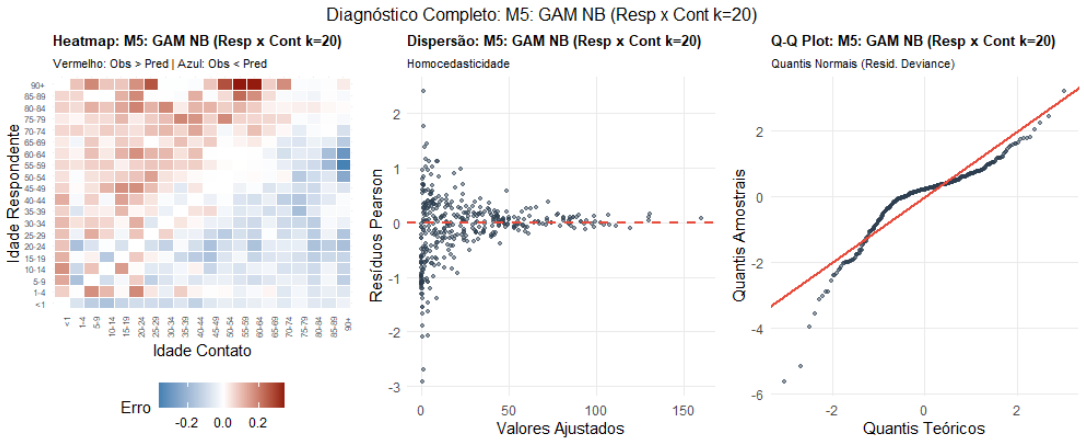
A análise dos resíduos para os modelos Binomial Negativa revelou nuances críticas sobre a interação entre a complexidade da base (k) e a estimação do parâmetro de dispersão (θ). Ao contrário da expectativa inicial de que a simples mudança de família resolveria a heterocedasticidade, observou-se que a flexibilidade da superfície é um pré-requisito para a estabilização da variância.

Figura 16 – Diagnóstico do Modelo Binomial Negativa M5 ($k = 15$).



Fonte: Elaboração da autora.

Figura 17 – Diagnóstico do Modelo Binomial Negativa M5 ($k = 20$).

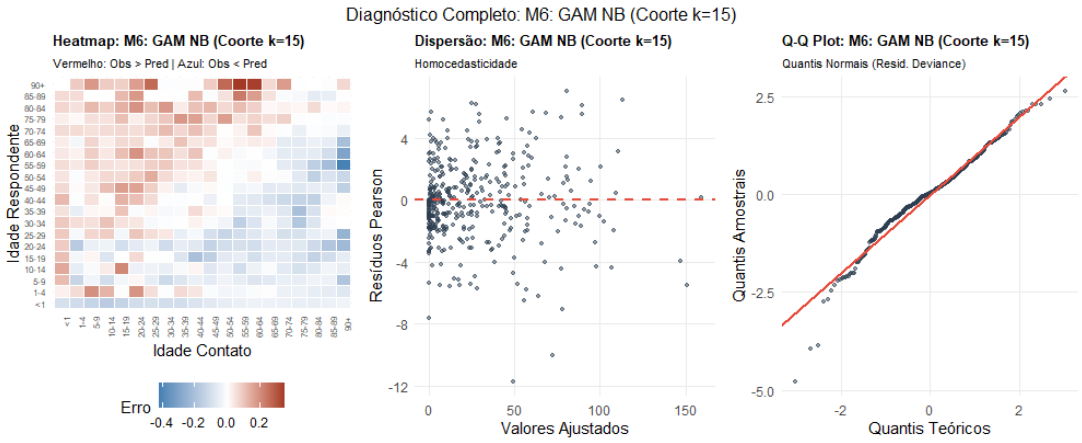


Fonte: Elaboração da autora.

A análise da Figura 16 (painel central) revela que o modelo M5 com $k = 15$ falhou em corrigir a heterocedasticidade. A dispersão dos resíduos mantém o formato característico do modelo Poisson, indicando que a variância continua crescendo linearmente com a média. Este comportamento ocorre porque a base de splines 15×15 na geometria cartesiana é excessivamente rígida para capturar o pico agudo de contatos na diagonal principal. Diante dessa falta de ajuste estrutural, o algoritmo de otimização não consegue estimar adequadamente o parâmetro de dispersão. O valor de θ divergiu para valores extremamente altos ($\theta \to \infty$), anulando o termo quadrático da variância ($Var(Y) \approx \mu + 0$) e fazendo com que a Binomial Negativa convergisse para um comportamento de Poisson.

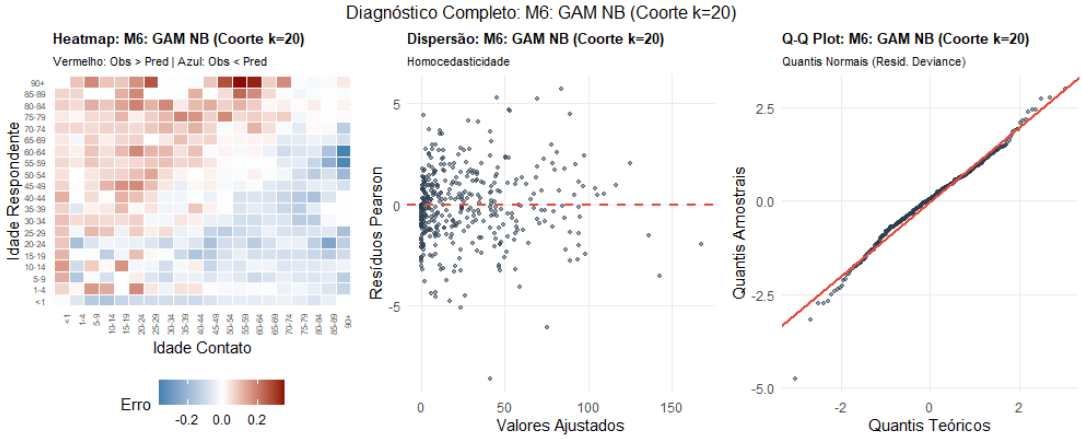
Ao aumentar a dimensão da base para $k = 20$ (Figura 17), observa-se uma correção drástica no gráfico de dispersão. A "nuvem" de resíduos torna-se uniforme, confirmando que a maior flexibilidade permitiu uma estimativa correta da média, possibilitando que o modelo encontrasse um θ finito e estabilizasse a variância. Contudo, a análise do Q-Q Plot (painel à direita da Figura 17) impõe cautela: embora a variância esteja estabilizada, os resíduos ainda apresentam desvios nas caudas. Isso sugere que, apesar de altamente aderente, o modelo M5 ainda sofre para acomodar os valores extremos da distribuição.

Figura 18 – Diagnóstico do Modelo Binomial Negativa M6 ($k = 15$).



Fonte: Elaboração da autora.

Figura 19 – Diagnóstico do Modelo Binomial Negativa M6 ($k = 20$).



Fonte: Elaboração da autora.

Em contraste, os modelos baseados em coorte (M6), apresentados nas Figuras 18 e 19, demonstraram comportamento superior nos Q-Q Plots. A aderência dos pontos à reta diagonal é quase perfeita, indicando que a geometria de coorte captura a estrutura dos dados de forma tão natural que os resíduos resultantes satisfazem rigorosamente as premissas distribucionais da Binomial Negativa, sem as caudas pesadas observadas no modelo cartesiano.

Análise Visual dos Padrões de Contato e Resíduos

Para além das métricas numéricas, é fundamental verificar se as matrizes geradas fazem sentido biológico. Nesta etapa, comparamos o que os modelos "enxergam" (a matriz suavizada) e onde eles "erram" (o mapa de calor dos resíduos), conectando esses achados

com os padrões clássicos estabelecidos na literatura, como o estudo POLYMOD Mossong et al. (2008).

1. O Modelo Padrão (M5): Ajustando o Foco da Imagem

O Modelo M5 assume que o contato depende da idade do respondente e a idade do contato relatado (eixos cartesianos). Na Figura 12 ($k = 15$), a matriz parece "desfocada" e a diagonal principal aparece alargada. Nos resíduos (Figura 16), observam-se grandes "manchas" coloridas, indicando que o modelo subestima sistematicamente faixas etárias inteiras. Ao aumentar a flexibilidade (Figura 13), a matriz ganha foco. Ao analisar a diagonal principal e as diagonais secundárias, notamos que o modelo captura a assortatividade intensa típica de crianças e adolescentes e o contato intergeracional (pais e filhos), essencial para modelar a transmissão domiciliar. Nos resíduos (Figura 17), as manchas sistemáticas desaparecem, restando apenas um "chiado" aleatório (ruído branco), indicando que o modelo capturou a estrutura latente.

2. O Modelo de Coorte (M6): Artefatos e Sobreajuste

O Modelo M6 tenta alinhar a suavização pela diferença de idade (d_{ij}), seguindo a lógica de que envelhecemos em coortes. As matrizes do M6 (Figuras 14 e 15) apresentam linhas diagonais muito marcadas, parecendo "estrias". A estrutura estriada observada viola a hipótese de continuidade da superfície de contato. Não há justificativa teórica para que a taxa de interação sofra descontinuidades severas baseadas em diferenças marginais de idade (ex: 1 ou 2 anos). Tais artefatos sugerem que a geometria de coorte, quando excessivamente flexível, impõe uma discretização artificial a um fenômeno que é intrinsecamente fluido.

Os mapas de resíduos do M6 (Figuras 18 e 19) são visualmente "perfeitos", apresentando cores muito claras. De forma didática, é como se o modelo tivesse "decorado a resposta" da amostra em vez de aprender o padrão populacional. Ele absorveu tanto o sinal quanto o ruído, o que explica sua falha ao tentar prever dados novos (alto erro na validação cruzada).

Conclusão Visual:

Enquanto o M6 cria padrões artificiais (estrias) para zerar o erro, o **M5 ($k = 20$)** oferece o equilíbrio ideal: ele recupera a nitidez dos padrões de contato descritos na literatura (alta interação entre jovens e mistura entre gerações) sem incorporar estruturas artificiais que não existem na dinâmica real.

3.3.2 Análise Comparativa de Desempenho e Seleção do Modelo

A Tabela 3 apresenta o "juiz final" para a seleção do modelo, confrontando métricas de ajuste interno (In-Sample) com a capacidade de generalização (Out-of-Sample).

Tabela 3 – Comparativo Final: Modelos Binomial Negativa.

Modelo	Estrutura	k	EDF	Dev. Exp.	AIC	CV-RMSE
M5	Padrão (Resp × Cont)	15	190,3	94,6%	19.864	0,130
M5	**Padrão (Resp × Cont)**	**20**	**324,5**	**99,8%**	**15.808**	**0,079**
M6	Coorte (Resp × Diff)	15	166,6	96,6%	18.174	0,111
M6	Coorte (Resp × Diff)	20	236,6	98,2%	16.919	0,157

Fonte: Elaboração da autora.

Nota: CV-RMSE = Erro Quadrático Médio da Validação Cruzada (menor é melhor).

3.3.2.0.1 Análise de Ajuste Interno (Deviance e AIC):

O aumento da dimensão da base de $k = 15$ para $k = 20$ (elevando o número teórico de funções de base de 224 para 399) resultou em ganhos substanciais de ajuste para ambas as estruturas. O modelo M5 ($k = 20$) atingiu uma Deviance Explicada impressionante de 99,8%, com uma queda drástica no AIC (de 19.864 para 15.808), indicando um ajuste quase perfeito aos dados de treinamento.

3.3.2.0.2 Análise de Generalização (CV-RMSE):

A métrica de validação cruzada revelou um comportamento divergente crucial entre as estruturas:

- **Modelo Padrão (M5):** O aumento da complexidade melhorou a capacidade preditiva. O CV-RMSE caiu de 0,130 ($k = 15$) para 0,079 ($k = 20$). Isso demonstra que o modelo não está apenas memorizando os dados, mas capturando nuances reais da superfície de contato que bases menores não conseguiam resolver.
- **Modelo de Coorte (M6):** O aumento da complexidade causou sobreajuste (*overfitting*). Embora o ajuste interno tenha melhorado (Deviance foi a 98,2%), o erro fora da amostra aumentou significativamente, subindo de 0,111 ($k = 15$) para 0,157 ($k = 20$). Isso sugere que a imposição rígida da estrutura diagonal (d_{ij}), quando combinada com alta flexibilidade, força o modelo a ajustar ruídos locais específicos da amostra que não se generalizam para novos dados.

4 Conclusão

O presente estudo abordou o desafio de estimar padrões de contato social no Complexo de Manguinhos, superando limitações metodológicas inerentes a dados de inquéritos populacionais em áreas vulneráveis. A correção inicial do sub-relato, baseada no mínimo domiciliar, provou-se indispensável para recuperar a consistência lógica das interações intra-domiciliares, transformando uma base de dados ruidosa e inflacionada por zeros em um conjunto passível de modelagem estatística.

A exploração de diferentes técnicas de suavização permitiu identificar a natureza estocástica dos contatos sociais. A abordagem não-paramétrica inicial via LOESS (Modelos M1 e M2), embora útil para visualização exploratória, revelou-se insuficiente para inferência final. O diagnóstico de resíduos evidenciou uma severa heterocedasticidade, confirmando que tratar taxas de contato como variáveis contínuas com variância constante viola as propriedades fundamentais de dados de contagem.

A transição para os Modelos Aditivos Generalizados (GAM) expôs a inadequação da família Poisson (M3) diante da alta variabilidade dos dados. A detecção de uma sobredispersão significativa ($\phi \approx 16$) invalidou a premissa de igualdade entre média e variância. Embora a abordagem Quasi-Poisson (M4) tenha corrigido a inferência dos erros-padrão, ela resultou em uma penalização excessiva da rugosidade, produzindo superfícies "borradas" que perderam detalhes importantes da estrutura de contato. Essas limitações conduziram à adoção da distribuição Binomial Negativa, que teoricamente permite modelar a heterogeneidade da população através do parâmetro de dispersão θ.

Entretanto, a análise diagnóstica revelou que a simples mudança de distribuição, sem a flexibilidade adequada, não garante a correção do modelo. Observou-se que o modelo Binomial Negativa com base reduzida (M5, $k = 15$) sofreu de uma rigidez estrutural que impediu o ajuste ao pico da diagonal principal. Como consequência, o algoritmo falhou em estimar o parâmetro de dispersão, fazendo com que θ divergisse para valores extremamente altos. Isso provocou a degeneração do modelo para um comportamento funcionalmente idêntico ao de Poisson, falhando em estabilizar a variância e mantendo o padrão cônico nos resíduos.

Superada essa limitação pelo aumento da complexidade da base, a comparação final concentrou-se na estrutura geométrica dos preditores (Idade-Idade vs. Coorte) e na capacidade de generalização.

Embora o modelo de Coorte (M6) tenha apresentado resíduos com aderência quase perfeita à distribuição teórica nos Q-Q Plots, a análise integrada revelou que esse comportamento é sintomático de *overfitting*. Ao moldar-se excessivamente à geometria diagonal da amostra, o M6 gerou artefatos visuais (estrias) biologicamente implausíveis e perdeu ca-

pacidade de generalização, resultando em um erro de predição fora da amostra (CV-RMSE = 0,157) substancialmente superior ao do modelo padrão.

Em contrapartida, o modelo M5 (Binomial Negativa, $k = 20$) demonstrou ser a escolha mais robusta. Apesar dos ligeiros desvios nas caudas do Q-Q Plot, este modelo apresentou o menor erro global de validação cruzada (CV-RMSE = 0,079) e produziu superfícies de contato visualmente coerentes com a literatura, recuperando a assortatividade e as interações intergeracionais sem incorporar ruído. Isso indica que sua estrutura cartesiana, quando dotada de flexibilidade suficiente, captura o sinal fundamental do processo de contato social, garantindo parâmetros mais confiáveis para simulações epidemiológicas futuras. Portanto, o M5 ($k = 20$) é eleito o modelo final deste estudo.

Dessa forma, o trabalho não apenas entrega à comunidade científica uma matriz de contato validada e inédita para a realidade de favelas do Rio de Janeiro, mas também sublinha o valor estratégico de analisar padrões de interação em populações altamente vulneráveis. A correta representação dessas dinâmicas é decisiva: ao substituir pressupostos genéricos por parâmetros empíricos robustos e localizados, modelos epidemiológicos tornam-se capazes de projetar cenários muito mais fidedignos. Consequentemente, a aplicação destas matrizes fundamenta a elaboração de políticas públicas mais assertivas e equitativas. Tais ferramentas permitem desenhar intervenções — como estratégias de vacinação ou medidas de mitigação — com a precisão necessária para proteger comunidades historicamente expostas a riscos sanitários e, simultaneamente, controlar mecanismos de transmissão em redes de alta densidade que, se negligenciadas, poderiam atuar como vetores de superespalhamento, beneficiando assim a saúde pública como um todo.

Referências

COELHO, L. E. et al. Prevalence and predictors of anti-sars-cov-2 serology in a highly vulnerable population of rio de janeiro: A population-based serosurvey. *The Lancet Regional Health–Americas*, Elsevier, v. 15, 2022.

KASSTEELE, J. van de; EIJKEREN, J. van; WALLINGA, J. Efficient estimation of age-specific social contact rates between men and women. 2017.

LITVINOVA, M. et al. Epistorm-mix: Mapping social contact patterns for respiratory pathogen spread in the post-pandemic united states. *medRxiv*, Cold Spring Harbor Laboratory Press, p. 2025–11, 2025.

MANNA, A. et al. Generalized contact matrices allow integrating socioeconomic variables into epidemic models. *Science Advances*, American Association for the Advancement of Science, v. 10, n. 41, p. eadk4606, 2024.

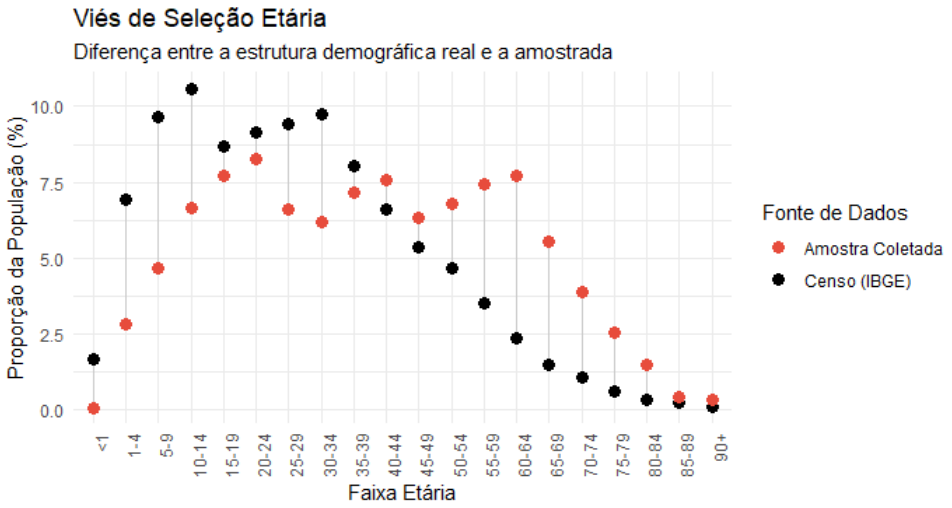
MELEGARO, A. et al. What types of contacts are important for the spread of infections? using contact survey data to explore european mixing patterns. *Epidemics*, Elsevier, v. 3, n. 3-4, p. 143–151, 2011.

MOSSONG, J. et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, Public Library of Science, v. 5, n. 3, p. e74, 2008.

VANDENDIJCK, Y. et al. Cohort-based smoothing methods for age-specific contact rates. *Biostatistics*, Oxford University Press, v. 25, n. 2, p. 521–540, 2024.

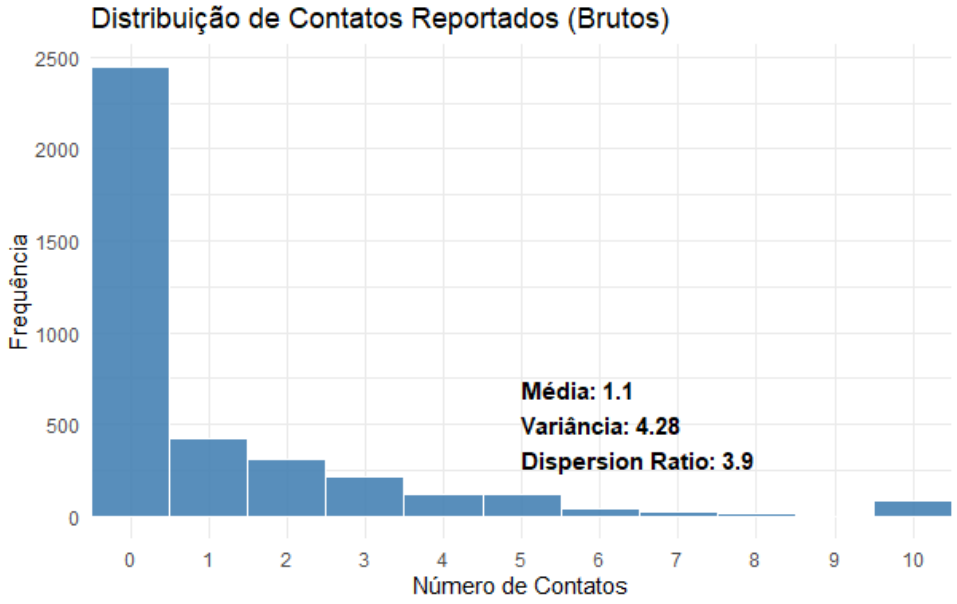
APÊNDICE A – Figuras Complementares

Figura 20 – Viés de Seleção Etária.



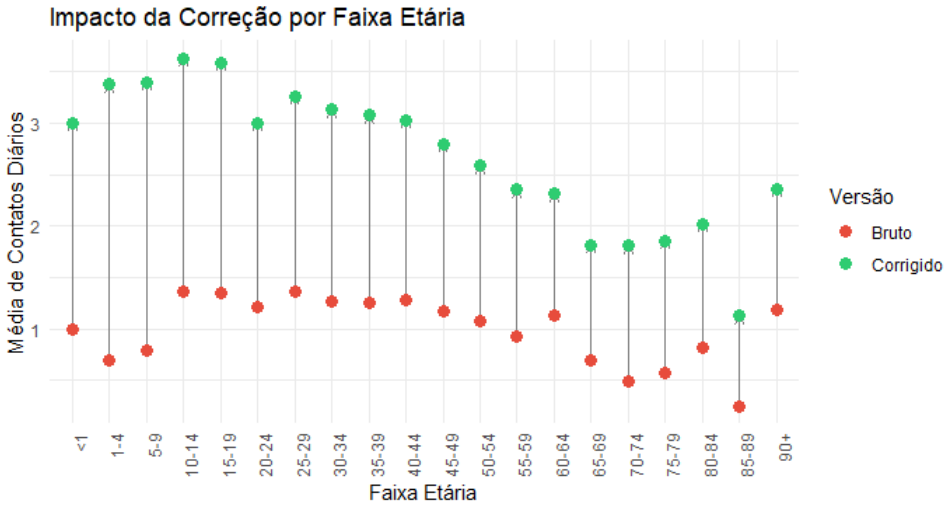
Fonte: Elaboração da autora.

Figura 21 – Distribuição de Contatos Reportados (Brutos).



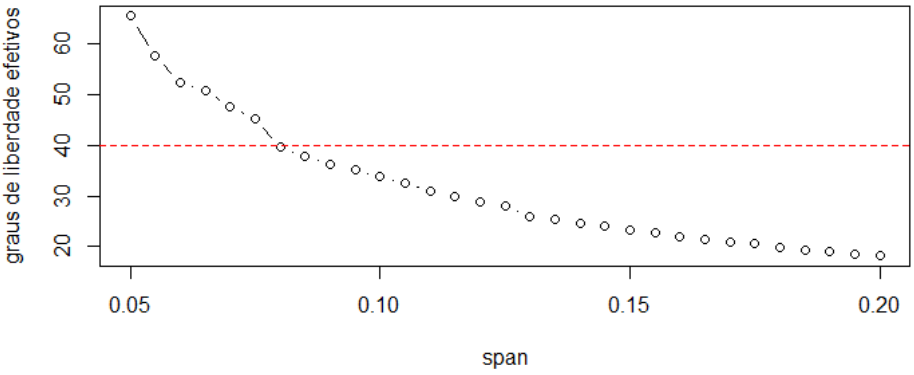
Fonte: Elaboração da autora.

Figura 22 – Impacto da Correção por Faixa Etária.



Fonte: Elaboração da autora.

Figura 23 – Calibração do parâmetro span.



Fonte: Elaboração da autora.