UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Estatística Aplicada

Ronaldo Marques Carvalho

Aplicação de modelos lineares generalizados mistos na classificação de surtos de dengue em unidades de saúde da cidade de Belo Horizonte

Belo Horizonte
2023

Ronaldo Marques Carvalho

Aplicação de modelos lineares generalizados mistos na classificação de surtos de dengue em unidades de saúde da cidade de Belo Horizonte

Monografia de especialização apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística Aplicada.

Orientador: Prof. Dr. Guilherme Lopes de Oliveira

Belo Horizonte
2023

2023, Ronaldo Marques Carvalho.
Todos os direitos reservados

Carvalho, Ronaldo Marques

C331a Aplicação de modelos lineares generalizados mistos na classificação de surtos de dengue em unidades de saúde da cidade de Belo Horizonte [recurso eletrônico] / Ronaldo Marques Carvalho – 2023.
1 recurso online 38 f. il., color.) : pdf.

Orientador: Guilherme Lopes de Oliveira
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 33-36

1. Estatística. 2. Modelos lineares (Estatística). 3. Dengue – Belo Horizonte (MG). 4. Dengue – Epidemiologia – Belo Horizonte (MG). I. Oliveira, Guilherme Lopes de.
II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 314ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE RONALDO MARQUES CARVALHO.

Aos doze dias do mês de dezembro de 2023, às 10:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Ronaldo Marques Carvalho**, intitulado: "Aplicação de modelos lineares generalizados mistos na previsão de surtos de dengue em unidades de saúde da cidade de Belo Horizonte", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Guilherme Lopes de Oliveira – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 12 de dezembro de 2023.

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 12/12/2023 [illegible]
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Lopes de Oliveira (Orientador)
DECOM/CEFET-MG

gov.br Documento assinado digitalmente
LEONARDO SOARES BASTOS
Data: 12/12/2023 [illegible]
Verifique em https://validar.iti.gov.br

Prof. Dr. Leonardo Soares Bastos
PROCC/Fiocruz

AGRADECIMENTOS

Meu sincero agradecimento ao professor orientador Guilherme Lopes de Oliveira que desde o primeiro contato se fez prontamente disponível para me auxiliar a encontrar os melhores caminhos para a execução dessa pesquisa.

Aos amigos da turma de Estatística Aplicada 2021 que em meio a todo o distanciamento tornaram o processo mais leve por meio das inúmeras horas em chamada de vídeo que passamos.

Por fim, agradeço a meu pai Ronaldo Abreu Carvalho, que desde o início da minha trajetória sempre me incentivou a seguir em frente. Essa especialização é mais um símbolo da nossa parceria de toda uma vida, ainda que não mais esteja comigo pessoalmente. Obrigado, pai.

RESUMO

O objetivo dessa pesquisa foi aplicar modelos lineares generalizados mistos (GLMM) para classificação de surtos de dengue em unidades de saúde na cidade de Belo Horizonte, Minas Gerais, Brasil, incluindo a informação de variáveis climáticas e efeitos espaço-temporais. Definiu-se um surto quando o total de dados notificados da doença foi igual ou excedeu o 95º percentil da série histórica de notificações mensais de dengue na unidade de saúde. O GLMM aplicado contou com dois efeitos aleatórios e quinze efeitos fixos. Os resultados demonstraram que variáveis climáticas como temperatura, precipitação e umidade tem efeito significativo na chance de ocorrência de surtos de dengue. Para algumas delas, estimou-se que a medida em um, dois ou até três meses anteriores podem refletir positivamente na chance de surte no mês corrente. No modelo final, apenas o efeito aleatório relacionado ao mês de ocorrência mostrou ser relevante para a explicação da variabilidade dos surtos de dengue. Para testar a qualidade das classificações foram utilizadas as medidas de acurácia, sensibilidade e especificidade, as quais retornaram capacidade preditiva acima de 91%.

Palavras-chave: epidemiologia; dengue; modelo linear generalizado misto; classificação de surto.

ABSTRACT

The aim of this study is to apply generalized linear mixed models (GLMMs) to forecast dengue outbreaks in health units in Belo Horizonte, Minas Gerais, Brazil, including information on climate variables and space-time effects. An outbreak is defined when the total number of dengue reports equal to or exceeding the 95th percentile of the historical series of monthly dengue reports in the health unit. The applied GLMM include two random effects and fifteen fixed effects. Results showed that climate variables such as temperature, precipitation, and humidity have a significant effect on the odds of dengue outbreaks. For some of them, it was estimated that the measurement in one, two, or even three previous months could affect the chance of outbreak in the current month. In the final model, only the random effect related to the month of occurrence was shown to be relevant for explaining the variability of dengue outbreaks. To test the quality of the classifications, the accuracy, sensitivity, and specificity measures were used, which returned predictive capacity above 91%.

Keywords: dengue; epidemiology; generalized linear mixed model; outbreak classification.

LISTA DE FIGURAS

Figura 1 – *Boxplots* das variáveis climáticas 26

Figura 2 – Curva de sensibilidade, especificidade e ponto de corte de acordo com o modelo GLMM ajustado 30

LISTA DE QUADROS E TABELAS

Quadro 1 – Variáveis da tabela de dados DENGUE_CLIMA 15

Quadro 2 – Medidas de avaliação baseadas na matriz de confusão 20

Tabela 1 – Exemplo da definição de surtos de acordo com o percentil 90 19

Tabela 2 – Matriz de confusão 20

Tabela 3 – Distribuição de surtos de acordo com o percentil considerado 22

Tabela 4 – Descrição e métricas de avaliação dos modelos testados 23

Tabela 5 – Distribuição dos surtos por mês e ano 24

Tabela 6 – Distribuição dos surtos por região 25

Tabela 7 – Análise exploratória das variáveis climáticas 26

Tabela 8 - Estimativas dos efeitos aleatórios do modelo 28

Tabela 9 – Resumo do modelo GLMM final ajustado 29

Tabela 10 – Matriz de confusão dos surtos de dengue de acordo com o modelo GLMM ajustado 31

SUMÁRIO

1 INTRODUÇÃO 11

2 MATERIAIS E MÉTODOS 13

2.1 Coleta, Limpeza e Seleção dos Dados 13

2.2 Definição de surtos 16

2.3 Modelo Linear Generalizado Misto (GLMM) 17

2.3.1 Definição de variável resposta Surto 18

2.3.2 Medidas de Avaliação da Capacidade Preditiva do Modelo 19

2.4 *Softwares* 20

3 RESULTADOS E DISCUSSÃO 22

3.1 Determinação do percentil de surto e variáveis para o modelo final 22

3.2 Distribuição dos surtos 24

3.3 Análise exploratória das variáveis climáticas 25

3.5 Predição de surtos 30

5 CONCLUSÃO 32

REFERÊNCIAS 33

ANEXOS 37

Anexo 1 – Função de Haversine 37

Anexo 2 - Alocar a estação climática às unidades de saúde 37

1 INTRODUÇÃO

A dengue é uma das doenças virais transmitida por mosquito que mais rápido se espalhou por regiões de clima tropical, tendo casos reportados em todos os Estados membro da Organização Mundial de Saúde (OMS) (OMS, 2022). O vetor da doença é a fêmea do mosquito *Aedes aegypti* que também é vetor da Chikungunya, Zika vírus e Febre Amarela. Estima-se que em 2017 mais de 100 milhões de pessoas foram infectadas em todo o mundo (ZENG *et al.,* 2021). Um levantamento da OMS estima que em 2019 houve a maior quantidade de casos reportados na história e, pela primeira vez, foram identificados casos no Afeganistão (OMS, 2022).

De acordo com Osanai *et al.* (1983), o primeiro registro de epidemia de dengue documentada clínica e laboratorialmente no Brasil ocorreu em Roraima em 1982, com mais de 10 mil casos confirmados. Após isso, novas epidemias foram registradas apenas em 1986 no Rio de Janeiro e posteriormente em todo o território brasileiro (SIQUEIRA *et al*., 2005; TEIXEIRA *et al*., 2013). Além da introdução de diferentes sorotipos da doença, o período de 2015-2016 também foi marcado pela introdução das arboviroses Chikungunya e Zika vírus (Ministério da Saúde, 2019). Em Belo Horizonte, capital do estado de Minas Gerais, a dengue tem sido considerada endêmica desde sua introdução em 1996 e, até então, a cidade passou por epidemias nos anos de 1998, 2010, 2013, 2016 e 2019 (CAMPOS *et al.*, 2019; SES-MG, 2019).

Os sintomas da doença variam de assintomáticos à gripe severa e em casos incomuns pode causar complicações como sangramentos e incapacitação do funcionamento de órgãos, levando os infectados à morte (OMS, 2022). A taxa de letalidade da doença é de aproximadamente 20% na ausência de um diagnóstico tempestivo e falta de tratamento correto (CAPEDING *et al.,*2014; HALSTEAD, 2012). Surtos da doença sobrecarregam os sistemas públicos de saúde, impedem que pessoas infectadas consigam o tratamento adequado e, eventualmente, elevam as taxas de letalidade (HALSTEAD, 2002). Percebe-se então o valor em classificar o comportamento da doença e o tamanho dos surtos com

antecedência para que recursos possam ser alocados de maneira eficiente mitigando possíveis impactos da doença (CHEN *et al.,* 2015; LEUNG *et al.,* 2023).

De acordo com a literatura, doenças virais transmitidas por meio de vetores demonstram grau elevado de heterogeneidade espacial devido a fatores como a variação espacial do vetor, padrões climáticos e medidas de controle humano (BRINDLE *et al.,* 2023; OMS, 2022; MBOGO *et al.,* 2003; ACEVEDO *et al.,* 2015). As diferentes relações entre os três fatores produzem os diversos padrões de proliferação e transmissão da doença que influenciam o potencial de ocorrência de surtos (TEURLAI *et al.,* 2015).

Neste sentido, modelos preditivos são constantemente elaborados explorando os fatores que melhor representam o comportamento da dengue em diferentes regiões ao redor do globo. As variáveis exploradas nesses modelos, geralmente, envolvem clima (BENEDUM *et al.,* 2018; COLÓN-GONZÁLEZ *et al.*, 2021; ALEIXO *et al.,* 2022), efeitos demográficos e socioeconômicos (QURESHI, TABINDA e VEHRA, 2017; BETANZOS *et al*., 2018; CHANG *et al*., 2015), fatores que envolvem os vetores (ABURAS, CETINER e SARI, 2010; GHARBI *et al*., 2011). O consenso encontrado na literatura é que ainda não há modelo único capaz de classificar com níveis satisfatórios todos os diferentes cenários globais da doença (LEUNG *et al*., 2023).

Tendo em vista os prejuízos provocados à saúde e qualidade de vida dos cidadãos, a intrínseca relação entre a proliferação do vetor de transmissão com as condições climáticas e a necessidade de que modelos preditivos sejam adotados levando em consideração as relações únicas entre os fatores das regiões afetadas, o presente estudo tem como objetivo aplicar um modelo linear generalizado misto para classificação de surtos de dengue em unidades de saúde na cidade de Belo Horizonte, incluindo a informação de variáveis climáticas e efeitos espaço-temporais.

2 MATERIAIS E MÉTODOS

Para se executar a estimação das probabilidades de surto de dengue foi utilizada a classe de modelos lineares generalizados mistos (GLMMs) com a consideração de variáveis climáticas, temporais e geográficas nos efeitos fixos e aleatórios, buscando capturar efeitos hierárquicos presentes no comportamento de disseminação da doença na cidade estudada. Os dados utilizados foram obtidos a partir de diferentes bases de dados públicas. A Seção 2.1 traz detalhes sobre o processo de obtenção e manipulação dos dados e a descrição da metodologia do GLMM é apresentada na Seção 2.2.

2.1 Coleta, Limpeza e Seleção dos Dados

Os dados de notificações de dengue foram coletados da base de dados abertos do DATASUS[1] , que fornece tabelas de dados anuais em que cada observação é uma notificação. A pesquisa foi conduzida com dados de janeiro de 2014 a dezembro de 2019, totalizando sete anos completos de dados mensais. Os dados disponibilizados pelo DATASUS anteriores a 2014 não possuem registro das unidades de saúde e, portanto, não foi possível utilizar um período anterior a este ano. Além disso, optou-se por não utilizar dados após 2019 por acreditar que o período da pandemia da COVID-19 possa ter influenciado as notificações de novos casos.

Após uma primeira análise das variáveis disponíveis foram identificadas três de interesse: *DT_SIN_PRI* (Data do sintoma principal), *ID_MUNICIP* (Identificação do município [310620 – Belo Horizonte]) e *ID_UNIDADE* (Identificação da unidade de saúde que notificou o caso). Foram então extraídas as variáveis de interesse para cada um dos anos de análise e em seguida foram agrupados por mês e por unidade de saúde, produzindo as novas variáveis *MES* (1-12, mês de ocorrência das notificações), *ANO* (2014-2019, ano de ocorrência das notificações) e *NOTIFI* (contagem das notificações naquele respectivo mês e ano para cada unidade de saúde). A variável que define o surto foi criada e adicionada à base de dados, formando a tabela de dados *DENGUE*. Por se tratar

[1] https://datasus.saude.gov.br

de uma definição importante e controversa, a criação da variável identificadora de surto será melhor discutida separadamente na Seção 2.2.

Os dados climáticos foram extraídos da base de dados abertos do Instituto Nacional de Meteorologia[2] (INMET) que fornece uma série histórica detalhada com dados climáticos por hora de cada estação climática em atividade no território brasileiro. Após análise das variáveis disponíveis e verificação na literatura quanto às variáveis climáticas geralmente utilizadas para esse tipo de modelagem, foram identificadas as seguintes variáveis de interesse: *DATA* (Data de registro), *ESTACAO* (Estação climática de origem [A521 ou F501] em atividade na cidade de Belo Horizonte), *PRECIPITACAO_TOTAL* (Total de chuvas [mm]), *TEMP_AR* (Temperatura do Ar [cº]), UMIDADE_REL_AR (Umidade relativa do ar [g/m³]) e *VENTO_RAJ_MAX* (Rajada de Vento Máxima [km/h]).

O teste paramétrico t-Student foi utilizado para verificar a diferença de médias entre os dados coletados por cada uma das estações climáticas e evidenciou diferenças estatisticamente significantes para todas as variáveis climáticas de interesse. Portanto, existe evidência suficiente para justificar a separação dos dados climáticos por estação e, com isto, as informações climáticas a serem utilizadas para uma determinada unidade de saúde serão aquelas registradas pela estação climática mais próxima a ela. As variáveis foram então agrupadas para refletir as médias mensais em cada uma das duas estações climáticas, dando origem à variável *MES* e *ANO* (idênticas às variáveis de mesmo nome criadas na base DENGUE), formando tabela de dados CLIMA.

Para possibilitar o correto agrupamento dos dados de notificações e climáticos foi elaborada uma terceira base, utilizada para armazenar as variáveis chave. A tabela de dados CHAVE possui as variáveis *ID_UNIDADE*, *CEP*, *LAT*, *LON* e *ESTACAO*. O processo de preenchimento dos dados ocorreu em três etapas. Na primeira etapa a variável *CEP* foi preenchida por meio da consulta dos identificadores das unidades de saúde no Cadastro Nacional de Estabelecimentos de Saúde (CNES). Na segunda etapa foram extraídos os dados de latitude e longitude das unidades de saúde e preencher as variáveis *LAT* e *LON*, respectivamente. As coordenadas geográficas das estações climáticas estavam disponíveis na base de dados do INMET e foram armazenadas em

[2] https://portal.inmet.gov.br/

um objeto separado. Por fim, na última etapa, foi criada uma função capaz de calcular a distância entre dois pontos por meio de suas respectivas coordenadas geográficas. Além disso, para cada unidade de saúde a função preencheu a variável *ESTACAO* com a identificação da estação climática mais próxima.

Após agrupar as variáveis, dando origem a tabela de dados DENGUE_CLIMA, optou-se por adicionar defasagens (*lags*) de três meses buscando melhor capturar os efeitos temporais das variáveis climáticas e de notificações, já que o mosquito precisa, além de tempo para se proliferar, de pessoas contaminadas para que a doença se espalhe. A escolha da quantidade de *lags* foi respaldada pela literatura (ALEIXO *et al.,* 2022; ZAW *et al.,* 2023) que ao testar diferentes defasagens, geralmente, apontou maior relevância nos *lags* entre 1 até 3 meses para variáveis climáticas.

Após a remoção das variáveis com observações faltantes e a adição dos *lags*, também foram adicionadas duas outras variáveis para fins de modelagem temporal: *TRIMESTRE* e *SEMESTE*. Todo o processo de manipulação dos dados deu origem à tabela DENGUE_CLIMA utilizada para modelagem, a qual conta com 10.480 observações e 34 variáveis conforme resumo apresentado no Quadro 1.

Quadro 1 – Variáveis da tabela de dados DENGUE_CLIMA

Variável	Descrição
ID_UNIDADE	Identificador das unidades de saúde. 299 no total
ESTACAO	Estação climática A521 ou F501
MES	1-12 mês do ano
ANO	2014-2019
TRIMESTRE	1-4 trimestres do ano
SEMESTRE	1-2 semestres do ano
NOTIFI	Contagem de notificações mensais
LAG_1_NOTIFI	Lag de 1 mês nas notificações mensais
LAG_2_NOTIFI	Lag de 2 meses nas notificações mensais
LAG_3_NOTIFI	Lag de 3 meses nas notificações mensais
PRECIP	Média de chuvas mensal em mm
LAG_1_PRECIP	Lag de 1 mês na média de chuva mensal em mm
LAG_2_PRECIP	Lag de 2 meses na média de chuva mensal em mm
LAG_3_PRECIP	Lag de 3 meses na média de chuva mensal em mm
TEMP	Temperatura média mensal em cº
LAG_1_TEMP	Lag de 1 mês na temperatura média mensal em cº
LAG_2_TEMP	Lag de 2 meses na temperatura média mensal em cº
LAG_3_TEMP	Lag de 3 meses na temperatura média mensal em cº
UMIDADE	Umidade média mensal em g/m³

LAG_1_UMIDADE	Lag de 1 mês na umidade média mensal em g/m³
LAG_2_UMIDADE	Lag de 2 meses na umidade média mensal em g/m³
LAG_3_UMIDADE	Lag de 3 meses na umidade média mensal em g/m³
VENTO	Média da velocidade média máxima mensal em km/h
LAG_1_VENTO	Lag de 1 mês na média da velocidade média máxima mensal em km/h
LAG_2_VENTO	Lag de 2 meses na média da velocidade média máxima mensal em km/h
LAG_3_VENTO	Lag de 3 meses na média da velocidade média máxima mensal em km/h
PERC80	Percentil 80 de notificações mensais por unidade de saúde
SURTO80	Variável dummy para ocorrência de surtos
PERC85	Percentil 85 de notificações mensais por unidade de saúde
SURTO85	Variável dummy para ocorrência de surtos
PERC90	Percentil 90 de notificações mensais por unidade de saúde
SURTO90	Variável dummy para ocorrência de surtos
PERC95	Percentil 95 de notificações mensais por unidade de saúde
SURTO95	Variável dummy para ocorrência de surtos

Fonte: Elaborado pelo autor.

2.2 Definição de surtos

Considerando todo o processo de manipulação dos dados até a versão final na qual os diferentes modelos GLMM foram testados, a definição do que seria considerado um surto de dengue foi uma das etapas mais relevantes do processo. Buscando encontrar respaldo na literatura quanto à definição de surto, o estudo de Brady *et al.* (2015) destacou-se por testar 102 definições para o termo baseadas em diferentes métodos com dados brasileiros.

Entre os principais resultados, os autores identificaram que: (i) as definições de surto são diferentes, portanto, produzem resultados discrepantes quando aplicadas aos mesmos dados o que impossibilita a comparação destes; (ii) ainda que as definições de surto sejam padronizadas, as características dos surtos são inconsistentes entre regiões; e (iii) o momento de início do surto varia significativamente de acordo com a definição adotada. As considerações finais do estudo apontam que ainda não há consenso entre qual definição é a mais adequada para definir um surto de dengue (Brady *et al*., 2015).

Neste sentido, no presente estudo optou-se pela definição adotada por Aleixo *et al.* (2022), que definem surtos médios e agudos como sendo o instante em que o total de casos iguala ou supera um certo percentil da série histórica das unidades de saúde. O processo para identificar o que seria um surto ou não, envolveu o teste de diferentes

percentis (80, 85, 90 e 95) em diferentes GLMM detalhados no item 3.1 da seção de resultados e discussão.

2.3 Modelo Linear Generalizado Misto (GLMM)

Buscando modelar as relações entre variáveis climáticas e temporais na quantidade de surtos de dengue em unidades de saúde na capital mineira, optou-se pela aplicação de um GLMM. Esta classe de modelos leva em consideração a existência de grupos hierárquicos para explicar os efeitos fixos, melhorando o poder de classificação dos modelos lineares generalizados (GLM) tradicionais. Modelos com essas características podem introduzir uma possível estrutura de correlação dos dados, ao flexibilizar a estrutura da matriz de variância-covariância (GALWAY, 2006; FATORETTO, 2016).

De modo geral, um GLMM pode ser representado pela seguinte equação

$$g(\mu_i) = \boldsymbol{X_i'\beta} + \boldsymbol{Z_i'u} + \boldsymbol{\varepsilon_i} \quad (1)$$

Em que $g(.)$ representa a função de ligação relacionada à média da variável resposta; X é uma matriz N x p de p variáveis explicativas; β é um vetor coluna p x 1 dos coeficientes de regressão de efeitos fixos; Z é uma matriz N x qJ para q efeitos aleatórios e J grupos; u é um vetor qJ x 1 de efeitos aleatórios para J grupos; e ε é um vetor coluna N x 1 dos erros que captaria efeitos não explicados pelos demais termos do modelo. Os efeitos aleatórios contidos no vetor $\boldsymbol{u}$ do preditor linear devem ser independentes e identicamente distribuídos. Pressupõe-se que os efeitos e erros aleatórios seguem a distribuição normal com média zero (HARTZEL, AGRESTI E CAFFO 2001).

Autores como Spyrides *et al.* (2007) consideram que em situações em que os dados são coletados repetidamente em unidades temporais e/ou geográficas, faz-se necessário considerar o comportamento entre e intra indivíduos, de modo a capturar o comportamento de cada unidade como um subgrupo geoespacial/temporal com características individuais.

Serão explorados modelos com efeitos aleatórios que medem a variabilidade das estimações por unidade de saúde e mês. Dessa forma, o modelo pretende incluir a informação de variáveis climáticas e efeitos espaço-temporais. Têm-se como um dos pressupostos adotados a inclusão da variável unidade de saúde (ID_UNIDADE) como parte dos efeitos aleatórios de todos os modelos que foram testados, sendo a inclusão dos demais efeitos analisados com base em métrica de qualidade do ajuste.

As estimativas da variância dos efeitos aleatórios em um modelo linear generalizado misto (GLMM) são importantes para entender a contribuição da variabilidade não explicada pelos termos fixos do modelo. Essas estimativas são essenciais para avaliar a relevância dos efeitos aleatórios e a adequação do modelo para lidar com a estrutura de grupos nos dados.

Já os efeitos fixos explorados na análise correspondem ao ano, notificações de dengue defasadas no tempo e variáveis climáticas defasadas no tempo. A defasagem busca capturar efeitos temporais passados em surtos presentes. Foram consideradas defasagens de até, no máximo, três meses anteriores, de forma que um certo nível de defasagem (por exemplo, três meses atrás) só seria mantido se o anterior (no exemplo, dois meses atrás) necessariamente estivesse no modelo.

2.3.1 Definição de variável resposta Surto

No contexto deste trabalho, a variável resposta binária Y é o surto de dengue observados por mês em cada uma das unidades de saúde ao longo de seis anos. No método GLMM os parâmetros são estimados com base no método de estimação via máxima verossimilhança, assumindo, neste caso, que a distribuição condicional da resposta Y é uma distribuição de Bernoulli com média π_i, ou seja, P($Y_i = 1) = \pi_i$. A função de ligação $g(\mu_i)$ considerada é a *logit*, caso canônico associado à distribuição de Bernoulli na classe dos GLM.

Neste caso, a interpretação dos parâmetros dos efeitos fixos é dada com base nos logitos gerados. A exponencial dos parâmetros estimados corresponde às razões de chances

(do inglês, *odds ratio*) de transmissão de LV a partir de cada fator que compõe o modelo selecionado. A razão de chance corresponde à razão das chances de ocorrência de um evento entre dois níveis da variável explicativa. Estas razões de chances, quando a variável resposta é ordinal, podem ser calculadas de várias maneiras constituindo os tipos: local, local-global e global. Os modelos logitos utilizam a forma local-global (FATORETTO, 2017), ou seja, são calculados de forma cumulativa.

A Tabela 1 demonstra por meio de um exemplo ilustrativo como foram elaboradas as variáveis de surto por unidade de saúde e mês. Basicamente, os percentis de cada unidade de saúde em cada mês foram utilizados para definir por meio da variável *dummy* a ocorrência de um evento (surto) ou não-evento (não surto). No exemplo, são mostrados os casos de apenas três unidades de saúde nos três primeiros meses do ano. A variável SURTO90 foi preenchida com o valor 1 sempre que o total de casos notificados ultrapassou o percentil calculado para aquela unidade de saúde.

Tabela 1 – Exemplo da definição de surtos de acordo com o percentil 90

Id_unidade	Mes	Nº Notificações	Perc90	SURTO90
1	jan	55	68	0
1	fev	60	68	0
1	mar	70	68	1
2	jan	98	95,4	1
2	fev	85	95,4	0
2	mar	20	95,4	0
3	jan	160	198	0
3	fev	200	198	1
3	mar	190	198	0

Fonte: Elaborado pelo autor.

2.3.2 Medidas de Avaliação da Capacidade Preditiva do Modelo

Para avaliar a qualidade da classificação do GLMM com resposta binária, têm-se a análise por meio da matriz de confusão que faz o confronto entre os resultados reais e classificados pelo modelo em: VP (Verdadeiros Positivos), FP (Falsos Positivos), FN (Falsos Negativos) e VN (Verdadeiros Negativos). A estrutura da matriz de confusão pode ser vista na Tabela 2.

Tabela 2 – Matriz de confusão

		REAL	
		Verdadeiro	Falso
PREVISTO	Verdadeiro	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Falso	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Elaborado pelo autor.

Os termos VP e VN sinalizam que os valores preditos são equivalentes aos observados, são os termos que quanto mais altos melhor é a capacidade preditiva do modelo. Os termos FP e FN apontam para a ocorrência de erros tipo I e erros tipo II, respectivamente. No contexto do presente estudo, os erros tipo I são aqueles em que o modelo reconhece a presença de um surto de casos de dengue em uma unidade de saúde quando na verdade não há. Os erros tipo II indicam a ausência de surto quando na verdade aquela unidade de saúde, de acordo com a classificação adotada, está passando por uma situação de surto de casos de dengue.

O Quadro 2 demonstra como são calculadas as medidas de avaliação dos resultados obtidos pela matriz de confusão utilizados neste estudo, são elas: acurácia, sensibilidade e especificidade. Neste estudo, compreende-se acurácia como à medida que indica a proporção de surtos e não surtos corretamente previstos com relação ao total de classificações. A sensibilidade é a capacidade do modelo de identificar os surtos corretamente considerando o total real de surtos. A especificidade mede a capacidade do modelo de identificar os não surtos corretamente considerando o total real de não surtos. Para todas estas métricas, quanto maior o seu valor, melhor.

Quadro 2 – Medidas de avaliação baseadas na matriz de confusão

Medida	Fórmula
Acurácia	$(VP + VN) / (VP + VN + FN + FP)$
Sensibilidade	$VP / (VP + FN)$
Especificidade	$VN / (VN + FP)$

Fonte: Elaborado pelo autor.

2.4 *Softwares*

Todas as estimativas foram elaboradas através da função *glmmTMB* (BROOKS *et al*., 2017) do pacote de mesmo nome no *software* R (TEAM R, 2010). Os gráficos foram

elaborados através da função *ggplotly* do pacote *ggplot2* (WICKHAM, 2016). A função Haversine, utilizada para encontrar as coordenadas geográficas das unidades de saúde e a função para encontrar a estação climática mais próxima da unidade de saúde foram implementadas manualmente e podem ser vistas nos Anexos 1 e 2 deste trabalho.

3 RESULTADOS E DISCUSSÃO

3.1 Determinação do percentil de surto e variáveis para o modelo final

A variável resposta de interesse é binária, correspondendo à identificação da ocorrência ou não de surto de dengue. Destaca-se que o “surto” é um evento extremamente raro e, portanto, tem-se um desbalanceamento nas categorias da variável resposta. A Tabela 3 evidencia a distribuição de eventos e não eventos gerados por meio dos percentis testados.

Tabela 3 – Distribuição de surtos de acordo com o percentil considerado

	Percentil 80	%	Percentil 85	%	Percentil 90	%	Percentil 95	%
Evento	2.271	22%	1.739	17%	1.218	12%	708	7%
Não-evento	8.209	78%	8.741	83%	9.262	88%	9.772	93%

Fonte: Elaborado pelo autor.

Os resultados demonstrados na Tabela 3, evidenciam que quanto mais restrito é o percentil definido, menor é a quantidade de eventos (surtos) identificados, o que surte direto impacto na qualidade de ajuste do modelo devido à variável ser altamente desbalanceada. É possível observar o valor máximo de eventos da tabela estando no percentil 80 com aproximadamente 22%, enquanto a menor quantidade de eventos é observada no percentil 95 com aproximadamente 7%. Além disso, é importante ressaltar que o uso de percentil mais elevado que os utilizados resultam em uma grande quantidade de unidades de saúde sem nenhum surto em um ou mais períodos do estudo, o que dificulta o processo de modelagem e estimação dos efeitos fixos e aleatórios.

Os diferentes modelos testados podem ser observados na Tabela 4. Ela detalha o percentil utilizado, os efeitos aleatórios e efeitos fixos incluídos, além das métricas de qualidade de ajuste AIC, BIC e logLik, partindo de um modelo GLMM nulo, sem efeitos fixos e apenas com efeito aleatório de unidade de saúde, até um modelo completo com efeitos fixos e aleatórios.

Tabela 4 – Descrição e métricas de avaliação dos modelos testados

Modelo	Efeito Aleatório Unidade de Saúde	Efeito Aleatório Região	Efeito Aleatório Mês	Efeitos Fixos	AIC	BIC	logLik
Nulo 80	SIM	NÃO	NÃO	-	10.965,4	10.979,9	-5.480,7
Nulo 85	SIM	NÃO	NÃO	-	9.428,5	9.443,0	-4.712,2
Nulo 90	SIM	NÃO	NÃO	-	7.540,7	7.555,2	-3.768,3
Nulo 95	SIM	NÃO	NÃO	-	5.191,5	5.206,0	-2.593,8
Nulo 95	SIM	SIM	NÃO	-	5.193,5	5.215,3	-2.593,8
Nulo 95	SIM	NÃO	SIM	-	4.181,0	4.202,8	-2.087,5
Nulo 95	SIM	SIM	SIM	-	4.183,0	4.212,1	-2.087,5
GLMM 95	SIM	NÃO	SIM	Ano + Notificações (lags 1-3) + Precipitação (lag 2) + Temperatura (lags 1-3) + Umidade (lags 1-2) + Vento (lag 1)	2.400,7	2.531,3	-1.182,4

Fonte: Elaborado pelo autor.

A Tabela 4 demonstra a melhoria nas medidas de ajuste dos modelos nulos com efeitos aleatórios de unidade de saúde à medida que se alteram os percentis de 80 para 95. Em sequência, observam-se os efeitos aleatórios de região e mês adicionados ao modelo nulo com percentil 95 em diferentes combinações. Por fim, o modelo com efeitos aleatórios e fixos com percentil 95 foi selecionado para aplicação na classificação de surtos deste estudo por resultar nos menores valores de AIC e BIC e maiores valores para o logLik, indicando melhor qualidade no ajuste do modelo em comparação com os demais testados. Os efeitos fixos foram escolhidos de acordo com o nível de significância de 95% usando o critério *backward*. Ou seja, partiu-se do modelo com todos as variáveis explicativas (efeitos fixos) disponíveis, retirando-as uma a uma com base na menor significância.

3.2 Distribuição dos surtos

A Tabela 4 destaca a distribuição de surtos por ano e por mês com base nos dados definidos para análise. Entre o período de 2014 a 2019, foi identificado um total de 383.817 notificações de casos de dengue no município de Belo Horizonte, cujos picos de surtos foram registrados em 2016 e 2019, 54,66% e 42,80%, respectivamente. A análise dos surtos por mês exibida na Tabela 5 revela que aproximadamente 99% dos surtos mensais estão concentrados no primeiro semestre, entre os meses de janeiro a junho, sendo concentrados principalmente nos meses de março (25,99%), abril (27,40%) e maio (23,45%).

Tabela 5 – Distribuição dos surtos por mês e ano

Mês	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	Total (%)
Janeiro	0	0	12	1	1	0	14 (1,98%)
Fevereiro	0	0	118	0	0	1	119 (16,81%)
Março	0	0	169	1	0	14	184 (25,99%)
Abril	0	4	78	0	0	112	194 (27,40%)
Maio	1	5	7	0	0	153	166 (23,45%)
Junho	0	1	0	1	0	18	20 (2,82%)
Julho	0	0	0	0	0	1	1 (0,14%)
Agosto	0	0	1	0	0	1	2 (0,28%)
Setembro	0	0	2	1	0	0	3 (0,42%)
Outubro	0	0	0	0	1	2	3 (0,42%)
Novembro	0	0	0	1	0	1	2 (0,28%)
Dezembro	0	0	0	0	0	0	0 (0,00%)
TOTAL (%)	1 (0,14%)	10 (1,41%)	387 (54,66%)	5 (0,71%)	2 (0,28%)	303 (42,8%)	708 (100,0%)

Fonte: Elaborado pelo autor.

Ao observar a distribuição dos surtos por região na Tabela 6, é possível perceber que a região Centro-Sul da capital mineira obteve a maior quantidade de surtos durante o período analisado, seguida das regiões Nordeste, Barreiro, Oeste, Venda Nova, Norte, Noroeste, Leste e por fim, Pampulha com a menor quantidade.

Tabela 6 – Distribuição dos surtos por região

Região	Total
Barreiro	85 (12%)
Centro-Sul	137 (17,35%)
Leste	58 (8,19%)
Nordeste	86 (12,15%)
Noroeste	66 (9,32%)
Norte	72 (10,17%)
Oeste	77 (10,87%)
Pampulha	54 (7,63%)
Venda Nova	73 (10,31%)
TOTAL	**708 (100%)**

Fonte: Elaborado pelo autor.

3.3 Análise exploratória das variáveis climáticas

A análise exploratória apresentada na Tabela 7 e ilustrada na Figura 1 aponta as características que podem estar associadas aos surtos de dengue na cidade de Belo Horizonte durante o período estudado. Em termos medianos, nota-se que em períodos de surto a precipitação, temperatura e umidade estiveram um pouco mais altas, com menor intensidade de ventos.

Tabela 7 – Análise exploratória das variáveis climáticas

	Precipitação		Temperatura		Vento		Umidade	
	Surto	**Não Surto**	**Surto**	**Não Surto**	**Surto**	**Não Surto**	**Surto**	**Não Surto**
Mínimo	0	0	17,42	15,02	3,83	3,83	51,66	48,40
1º Quartil	0,06	0,02	21,37	19,87	4,23	4,86	68,04	62,29
Mediana	0,14	0,10	23,13	21,72	4,88	5,59	70,28	67,77
Média	0,15	0,16	22,53	21,54	5,63	6,38	69,27	66,04
3º Quartil	0,16	0,28	23,69	23,32	7,45	8,27	70,99	70,99
Máximo	0,5	0,53	25,07	25,75	10,33	10,96	80,60	80,60

Fonte: Elaborado pelo autor.

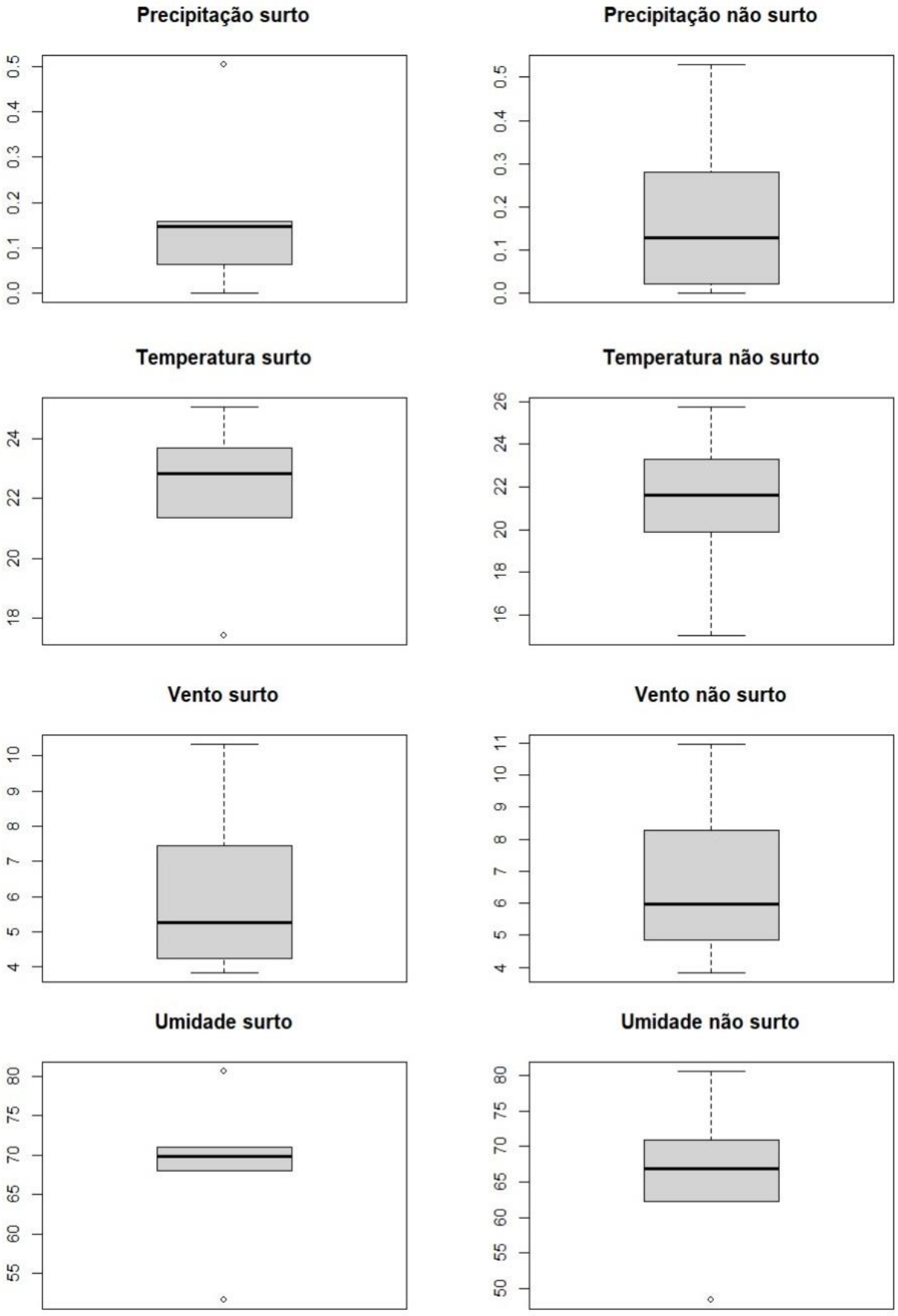
A média da precipitação durante surtos e não surtos foi de aproximadamente 0,15~0,16 mm. As diferenças mais relevantes são observadas no terceiro quartil que durante surtos foi de 0,16mm e durante não surtos foi de 0,28mm. Isto aponta que durante os surtos, 75% dos níveis de precipitação observados encontram-se em um intervalo de até 0,12mm menor do que o observado durante não surtos. A distribuição da precipitação evidencia que em períodos de surtos estão concentrados em um intervalo menos amplo do que em períodos em que não são registrados surtos, ou seja, há menor variabilidade.

Quanto à temperatura durante surtos e não surtos, observou-se uma diferença média de aproximadamente 1ºC, sendo a média durante surtos mais alta do que durante não surtos. A temperatura mínima durante surtos foi 2,4ºC maior do que durante não surtos, enquanto a mediana foi 1,41ºC superior. As demais análises apresentaram diferenças menores. Observou-se que a amplitude da temperatura – temperatura máxima - temperatura mínima – durante surtos foi de 7,65ºC enquanto durante não surtos foi de 10,73ºC, demonstrando que durante os eventos a temperatura é mais concentrada do que durante não eventos.

A velocidade média do vento durante surtos é 0,75 km/h menor do que durante não surtos. Ao analisar a mediana a diferença é similar, aproximadamente 0,71 km/h. O terceiro quartil aponta a maior diferença entre os dois grupos, 0,82 km/h.

A diferença média da umidade foi de aproximadamente 3,23 g/m³, enquanto a mediana foi de 2,51 g/m³. Em ambos os casos, as estatísticas descritivas dos surtos foram maiores do que durante não surtos. A amplitude dos dados durante surtos foi aproximadamente 28,94 g/m³ enquanto durante não surtos foi de 32,20 g/m³, novamente os dados durante surtos são mais concentrados do que durante não surtos.

Figura 1 – *Boxplots* das variáveis climáticas



Fonte: Elaborado pelo autor.

3.4 Modelagem

Conforme discutido na Seção 3.1, o GLMM aplicado utiliza como critério de definição de surtos quantidades de notificações por unidades de saúde iguais ou superiores a 95% do total de casos de toda a série histórica analisada, inclui dois efeitos aleatórios (unidade de saúde e mês) e quinze variáveis explicativas (ano de notificação e variáveis climáticas com lags de 1 a 3 meses).

Por meio da Tabela 8 é possível observar que a estimativa da variância do efeito aleatório referente ao mês apresenta um valor numericamente relevante, mas o mesmo não acontece para a variância do efeito aleatório relacionado às unidades de saúde. Quando a variância de um efeito aleatório é muito pequena, isso geralmente indica que a contribuição desse efeito aleatório para a variação total dos dados é insignificante ou quase nula. Em outras palavras, a variação entre os níveis ou grupos desse efeito aleatório é tão pequena que não faz uma diferença substancial na explicação da variação dos dados. Contudo, embora numericamente pequena, julga-se que a inclusão do efeito de área (unidade de saúde) é importante para retratar a presença de medidas repetidas em um mesmo local.

Tabela 8 - Estimativas dos efeitos aleatórios do modelo

Grupos	Quantidade	Variância	Desvio Padrão
ID_UNIDADE	299	1,585E-09	3,982E-05
MÊS	12	1,851	1,360

Fonte: elaborado pelo autor. Legenda: 1,585E-09=0,000000001585.

A Tabela 9 contém as estimativas e análise de significância para os efeitos fixos. O resumo do GLMM demonstra como os anos de 2016 e 2019 são mais críticos em termos de influência nas probabilidades de surtos de dengue, um resultado condizente com a distribuição de surtos evidenciado pela Tabela 5, enquanto os anos de 2015, 2017 e 2018 não apresentam efeito significativo na chance de ocorrência de surto quando comparados ao ano de referência, 2014. A *odds ratio* de ocorrência de surtos em 2016 ultrapassa 250% em comparação com o ano referência (2014), o que em 2019 foi de aproximadamente de 130%. As notificações defasadas apresentam significância

estatística para os três meses, sendo que de acordo com a *odds ratio* estimada, a cada notificação de dengue adicional em um determinado mês, aumenta-se em 0,5% a chance de surtos do mês seguinte, enquanto que com relação a dois e três meses atrás, um aumento de notificação reduz em 0,4% e 0,8% a chance de surto. O resultado pode ser relacionado ao ciclo de vida do mosquito e duração da infecção em indivíduos que é menor do que trinta dias. Isso também evidencia o comportamento sazonal da doença, indicando que não é esperada a manutenção de riscos elevados de surto durante muitos meses seguidos.

Tabela 9 – Resumo do modelo GLMM final ajustado

Variável	Coeficiente	Erro padrão	z valor	P valor	*Odds Ratio*
(Intercepto)	-33,39	5,113	-6,531	<0,001	-
Ano_2015	1,054	1,058	0,997	0,319	2,870
Ano_2016	5,577	1,018	5,478	<0,001	264,408
Ano_2017	1,933	1,105	1,749	0,802	6,911
Ano_2018	-0,534	1,238	-0,432	0,666	0,586
Ano_2019	4,866	1,012	4,807	<0,001	129,752
lag_1_Notificações	0,005	<0,001	7.707	<0,001	1,005
lag_2_Notificações	-0,004	<0,001	-3,307	<0,001	0,996
lag_3_Notificações	-0,008	<0,001	-3,591	<0,001	0,992
lag_2_Precipitação	3,112	0,912	3,412	<0,001	22,458
lag_1_Temperatura	0,320	0,013	2,439	0,0147	1,377
lag_2_Temperatura	-0,243	0,008	-3,119	<0,001	0,784
lag_3_Temperatura	0,213	0,005	4,566	<0,001	1,237
lag_1_Umidade	0,235	0,002	9,395	<0,001	1,265
lag_2_Umidade	0,056	0,002	2,292	0,022	1,057
lag_1_Vento	-0,210	0,008	-2,635	<0,001	0,811

Fonte: Elaborado pelo autor.

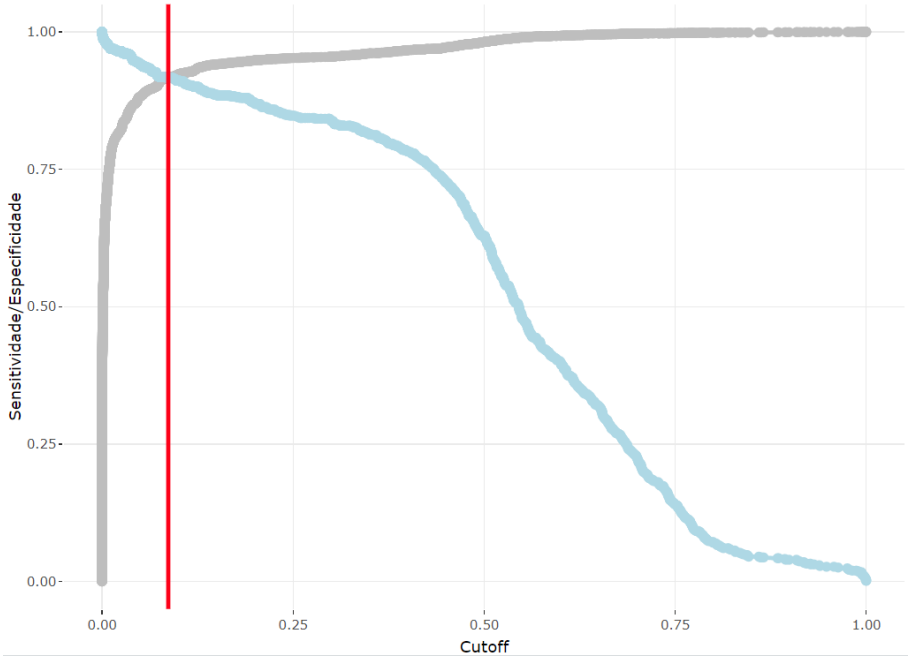
Ainda de acordo com a Tabela 9, a precipitação mostrou-se significante apenas com a defasagem de dois meses, apontando que um incremento unitário no nível de chuvas a dois meses atrás aumenta a chance de surtos de dengue em 22% no mês atual. As variáveis defasados de temperatura também apresentaram significância estatística e coeficientes positivos para as defasagens de um e três meses, indicando que aumentos na temperatura nesses períodos tendem a favorecer o aumento de 37,7% e 23,7% na chance surto de dengue, respectivamente. No entanto, o aumento da temperatura de dois meses atrás aponta para uma redução de aproximadamente 22% nas chances de surto no mês corrente. As variáveis de umidade apresentaram significância estatística,

indicando que aumentos na umidade (no mês atual e no mês anterior) influenciam positivamente na chance de surto de dengue. A variável vento apresentou o coeficiente negativo, indicando que um aumento no vento de um mês atrás reduz em 19% a quantidade de surtos do mês seguinte, sendo um efeito de difícil interpretação.

3.5 Predição de surtos

A predição de surtos foi realizada utilizando como utilizando como ponto de corte o valor de 0.088 que, aproximadamente, maximiza as medidas de classificação (especificidade, sensibilidade e acurácia) sendo demarcado em vermelho na Figura 2. Ela evidencia o valor de especificidade e sensibilidade em cada ponto de corte entre 0,0 e 1,0 por meio das curvas de cor cinza e azul respectivamente. Por meio do resultado obtido pela classificação com ponto de corte que maximiza as medidas de análise foi obtida a matriz de confusão representada na Tabela 10.

Figura 2 – Curva de sensibilidade, especificidade e ponto de corte de acordo com o modelo GLMM ajustado



Fonte: Elaborado pelo autor.

A matriz de confusão contida na Tabela 10 destaca os resultados preditos pelo modelo e oferece resultados que podem ser analisados por meio das medidas de análise das predições dos surtos de dengue, sendo elas: Acurácia (0,9184), Sensibilidade (0,9152) e Especificidade (0,9186). Por meio dos resultados citados é possível perceber que o modelo foi capaz de classificar corretamente acima de 91% dos surtos de acordo com todas as três medidas.

Tabela 10 – Matriz de confusão dos surtos de dengue de acordo com o modelo GLMM ajustado

	Verdadeiro	Falso
Verdadeiro	648	795
Falso	60	8977

Fonte: Elaborado pelo autor.

5 CONCLUSÃO

Para atender ao objetivo de desenvolver um modelo linear generalizado misto para classificação de surtos mensais de dengue em unidades de saúde na cidade de Belo Horizonte, incluindo a informação de variáveis climáticas e espaço-temporais, foi proposto um modelo com dois efeitos aleatórios (unidade de saúde e mês) e que inclui como efeitos fixos, além do ano de notificação, quatro variáveis climáticas defasadas de 1 a 3 meses passados. O modelo final escolhido com base em medidas usuais de qualidade de ajuste contou com um total de quinze variáveis explicativas.

Como não existe consenso quanto ao que define um surto de dengue na literatura, foi tratado como surto todas as notificações que eram iguais ou superiores a 95% do total de notificações da série histórica mensal da unidade de saúde. O percentil 95 foi escolhido após a observação da distribuição de surtos e do melhor ajuste dos modelos testados em diferentes percentis, a saber 80, 85, 90 e 95.

Os efeitos aleatórios foram testados em diferentes combinações, sendo possível observar nos efeitos do modelo uma organização hierárquica (unidades de saúde) e temporal (mês). A variância estimada do efeito aleatório relacionado à unidade de saúde foi muito pequena, indicando que este nível hierárquico pode não ser importante para a explicação das chances de ocorrência de surtos. Quanto aos efeitos fixos, foram selecionados aqueles que forneceram ao modelo melhores ajustes de AIC, BIC e *logLik* em comparação aos modelos nulos e demais modelos testados. A inserção de *lags* das variáveis temporais permitiu observar como o clima de até três meses passados impacta os surtos do mês seguinte. Para testar a qualidade das classificações foram utilizadas as medidas de acurácia, sensibilidade e especificidade que retornaram taxas acima de 91% de acerto para todas as três medidas, indicando boa capacidade preditiva do modelo.

O presente estudo possui como parte das limitações a não inclusão de variáveis como densidade demográfica dos bairros e disponibilidade de saneamento básico a fim de melhorar a qualidade dos ajustes. Uma melhor avaliação de possíveis efeitos da multicolinearidade e até mesmo a criação de grupos hierárquicos podem levar a um GLMM possa se ajustar melhor à realidade da capital mineira.

REFERÊNCIAS

ABURAS H, CETINER B, SARI M. **Dengue confirmed-cases prediction: A neural network model.** Expert Syst Appl. 2010; 37: 4256–4260.

ACEVEDO MA, PROSPER O, LOPIANO K, RUKTANONCHAI N, CAUGHLIN TT, MARTCHEVA M, et al. Spatial heterogeneity**, host movement and mosquito-borne disease transmission.** PLoS One. 2015; 10: e0127552. https://doi.org/10.1371/journal.pone.0127552.

ALEIXO, R., KON, F., ROCHA, R., CAMARGO, M. S., & CAMARGO, R. Y. DE. **Predicting dengue outbreaks with explainable machine learning**. Proceedings. 2022. doi:10.1109/CCGrid54584.2022.00114.

BENEDUM CM, SEIDAHMED OME, ELTAHIR EAB, MARKUZON N. **Statistical modeling of the effect of rainfall flushing on dengue transmission in Singapore**. PLoS Negl Trop Dis. 2018; 12: e0006935. https://doi.org/10.1371/journal.pntd.0006935.

BETANZOS-REYES AF, RODRIGUEZ MH, ROMERO-MARTÍNEZ M, SESMA-MEDRANO E, RANGEL-FLORES H, SANTOS-LUNA R. **Association of dengue fever with Aedes spp. abundance and climatological effects**. Salud Publica Mex. 2018; 60: 12–20. https://doi.org/10.21149/8141.

BRADY, OLIVER J. et al. **Dengue disease outbreak definitions are implicitly variable**. Epidemics, v. 11, p. 92-102, 2015.

BRINDLE, H. E., BASTOS, L. S., CHRISTLEY, R., CONTAMIN, L., DANG, L. H., ANH, D. D., ... & CHOISY, M. **The spatio-temporal distribution of acute encephalitis syndrome and its association with climate and landcover in Vietnam.** BMC Infectious Diseases, 23(1), 1-15. 2023.

BROOKS ME, KRISTENSEN K, VAN BENTHEM KJ, MAGNUSSON A, BERG CW, NIELSEN A, SKAUG HJ, MAECHLER M, BOLKER BM. "**glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling**." The R Journal, 9(2), 378–400. 2017. doi:10.32614/RJ-2017-066.

CAMPOS NBD, MORAIS MHF, CEOLIN APR ET AL. **Twenty-Two years of dengue fever (1996–2017): an epidemiological study in a Brazilian city.** Int J Environ Health Res 2019: 1–10.

CAPEDING MR, TRAN NH, HADINEGORO SRS, ISMAIL HI, CHOPITAYASUNONDH T, CHUA MN, et al. **Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial**. Lancet. 2014; 384: 1358–1365. https://doi.org/10.1016/S0140-6736(14)61060-6.

CENTRO DE CONTROLE E PREVENÇÃO DE DOENÇAS (CDC). **About Dengue: What You Need to Know**. Disponível em: <https://www.cdc.gov/dengue/about/index.html>. Acessado em: novembro de 2022.

CHANG F, TSENG Y, HSU P, CHEN C, LIAN I, CHAO D. **Re-assess Vector Indices Threshold as an Early Warning Tool for Predicting Dengue Epidemic in a Dengue Non-endemic Country**. PLoS Negl Trop Dis. 2015; 9: e0004043. https://doi.org/10.1371/journal.pntd.0004043.

CHEN HL, HSIAO WH, LEE HC, WU SC, CHENG JW. **Selection and Characterization of DNA Aptamers Targeting All Four Serotypes of Dengue Viruses**. PLoS One. 2015; 10: e0131240. https://doi.org/10.1371/journal.pone.0131240.

COLÓN-GONZÁLEZ, FELIPE J. et al. **Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles**. PLoS medicine, v. 18, n. 3, p. e1003542, 2021.

FATORETTO, M. B. **Modelos para dados categorizados ordinais com efeito aleatório: uma aplicação à análise sensorial**. Dissertação (Mestrado em Ciências) - Universidade de São Paulo, 2016.

GALWAY, N. W. **Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance**. London: 376p. 2006.

GHARBI M, QUENEL P, GUSTAVE J, CASSADOU S, RUCHE GL, GIRDARY L, et al. **Time series analysis of dengue incidence in Guadeloupe, French West Indies:**

forecasting models using climate variables as predictors. BMC Infect Dis. 2011; 11: 166. https://doi.org/10.1186/1471-2334-11-166.

HALSTEAD SB. **Dengue vaccine development: a 75% solution?** Lancet. 2012; 380: 1535–1536. https://doi.org/10.1016/S0140-6736(12)61510-4.

HALSTEAD SB. **Dengue.** Curr Opin Infect Dis. 2002; 15: 471–476. https://doi.org/10.1097/00001432-200210000-00003.

HARTZEL, J., AGRESTI, A.; CAFFO, B. **Multinomial logit random effects models**. Statistical Modelling, v.1, n.2, p81-102. 2001.

LEUNG, XING YU et al. **A systematic review of dengue outbreak prediction models: Current scenario and future directions**. PLOS Neglected Tropical Diseases, v. 17, n. 2, p. e0010631, 2023.

MBOGO CM, MWANGANGI JM, NZOVU J, GU W, YAN G, GUNTER JT, et al. **Spatial and temporal heterogeneity of Anopheles mosquitoes and Plasmodium falciparum transmission along the Kenyan coast**. Am J Trop Med Hyg. 2003; 68: 734–742.

MINISTÉRIO DA SAÚDE. **Boletim Epidemiológico**. Brasília: Secretaria de Vigilância em Saúde, v. 50, p. 1-154, 2019.

ORGANIZAÇÃO MUNDIAL DE SAÚDE (OMS). **Who fact sheets: Dengue and severe dengue**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>. Acessado em novembro de 2022.

OSANAI, CARLOS H. et al. **Surto de dengue em Boa Vista, Roraima Nota previa**. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 25, n. 1, p. 53-54, 1983.

QURESHI EMA, TABINDA AB, VEHRA S. **Predicting dengue outbreak in the metropolitan city Lahore, Pakistan, using dengue vector indices and selected climatological variables as predictors**. J Pak Med Assoc. 2017; 67: 416–421.

SECRETARIA DE ESTADO DE SAÚDE DE MINAS GERAIS (SES-MG). **Plano Estadual de Contingência Doenças Transmitidas pelo Aedes 2019–2020**. Governo de Minas Gerais; 2019. Disponível em: <https://www.saude.mg.gov.br/images/documentos/Plano%20de%20Conting%C3%AAncia%20Estadual%20Aedes%202019_2020_final.pdf>. Acessado em: novembro de 2022.

SIQUEIRA JB, MARTELLI CMT, COELHO GE, SIMPLICIO AC DA R, HATCH DL. **Dengue and dengue hemorrhagic fever, Brazil, 1981-2002**. Emerg Infect Dis 2005; 11: 48–53.

SPYRIDES, M.H.C.; STRUCHINER, C.J.; BARBOSA, M.T.S.; KAC, G. **Análise de Dados com Medidas Repetidas**. KAC, G.; SICHIERI, R.; GIGANTE, D.P. In: Epidemiologia Nutricional. Rio de Janeiro: Editora Fiocruz/Atheneu, 2007.

TEAM, R. DEVELOPEMENT CORE. "**R: A language and environment for statistical computing.**" (No Title) (2010).

TEIXEIRA, MARIA GLÓRIA; SIQUEIRA JR., JOÃO BOSCO; FERREIRA, GERMANO L.C.; BRICKS, LÚCIA; JOINT, GRAHAM. **Epidemiological trends of dengue disease in Brazil (2000-2010): a systematic literature search and analysis.** PLoS neglected tropical diseases, v.99 7, n. 12, p. e2520, jan. 2013.

TEURLAI M, MENK***È***S CE, CAVARERO V, DEGALLIER N, DESCLOUX E, GRANGEON JP, et al. **Socio-economic and Climate Factors Associated with Dengue Fever Spatial Heterogeneity: A Worked Example in New Caledonia**. PLoS Negl Trop Dis. 2015; 9: e0004211. https://doi.org/10.1371/journal.pntd.0004211.

ZAW, WIN et al. **Dengue in Myanmar: Spatiotemporal epidemiology, association with climate and short-term prediction.** PLOS Neglected Tropical Diseases, v. 17, n. 6.

ZENG Z, ZHAN J, CHEN L, CHEN H, CHENG S. **Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017**. EClinicalMedicine. 2021; 32: 100712. https://doi.org/10.1016/j.eclinm.2020.100712

WICKHAM H (2016). **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

ANEXOS

Anexo 1 – Função de Haversine

```
### isolando os dois pontos das estações automáticas
clima_bh_estacao <- select(clima_BH, cod_estacao,VL_LATITUDE, VL_LONGITUDE)
clima_bh_estacao <- distinct(clima_bh_estacao, cod_estacao, VL_LATITUDE, VL_LONGITUDE)

#Criar função para calcular a distancia de haversine
dist_haversine <- function(lat1, lon1, lat2, lon2, r = 6371) {
  dlat <- (lat2 - lat1) * pi / 180
  dlon <- (lon2 - lon1) * pi / 180
  a <- sin(dlat / 2) ^ 2 + cos(lat1 * pi / 180) * cos(lat2 * pi / 180) * sin(dlon / 2) ^ 2
  c <- 2 * asin(min(1, sqrt(a)))
  r * c
}
```

Anexo 2 - Alocar a estação climática às unidades de saúde

```
# função para encontrar a estação mais próxima
encontrar_estacao <- function(lat, lon, clima_bh_estacao) {
  if (!is.numeric(lat) || !is.numeric(lon)) {
    stop("lat and lon must be numeric")
  }

# calcula a distância entre a observação e cada estação
distancias <- apply(clima_bh_estacao[, c("VL_LATITUDE", "VL_LONGITUDE")], 1, function(x) dist_haversine(x[1], x[2], lat, lon))
```

```
# encontra a estação mais próxima
cod_estacao <- clima_bh_estacao$cod_estacao[which.min(distancias)]
  return(cod_estacao)
}

# converter lat and lon para tipo numérico
ceps$lat <- as.numeric(ceps$lat)
ceps$lon <- as.numeric(ceps$lon)

# adiciona a variável estação
ceps$estacao <- apply(ceps[, c("lat", "lon")], 1, function(x) encontrar_estacao(x[1], x[2], clima_bh_estacao))

ceps$estacao <- apply(ceps[, c("lat", "lon")], 1, function(x) encontrar_estacao(x[1], x[2], clima_bh_estacao))
```