



UNIVERSIDADE FEDERAL DE MINAS GERAIS

---

## Trabalho de Conclusão de Curso

---

*Aluno*  
Pedro Barbosa Bahia

*Orientadores*  
Frederico Coelho e Renato Assunção

Janeiro de 2025

# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>2</b>
<b>2</b>	<b>Pesquisa Bibliográfica</b>	<b>2</b>
2.1	Área de Estudo . . . . .	2
2.2	Ovitrampas . . . . .	2
2.3	Breve Histórico . . . . .	3
2.4	Estado da Arte . . . . .	5
<b>3</b>	<b>Descrição dos Dados</b>	<b>5</b>
3.1	Ovitrampas . . . . .	5
3.2	Variáveis exógenas . . . . .	9
3.2.1	Dados Metereológicos . . . . .	9
<b>4</b>	<b>Metodologia</b>	<b>11</b>
4.1	Seleção de entradas . . . . .	18
4.2	Modelos de Aprendizado de Máquina . . . . .	19
4.3	Métricas . . . . .	19
4.4	Seleção de Features . . . . .	20
4.5	Resíduos . . . . .	22
4.5.1	Classificação de Presença . . . . .	22
4.5.2	Régressão . . . . .	22
4.5.3	Cross Validation . . . . .	24
<b>5</b>	<b>Conclusão</b>	<b>24</b>
<b>6</b>	<b>Apêndice</b>	<b>24</b>
6.1	Tratamento de Valores Inconsistente . . . . .	24
6.2	Modelos iniciais . . . . .	24
<b>7</b>	<b>Bibliografia</b>	<b>25</b>

## 1 Introdução

Infecções arbovirais, como Dengue, Chikungunya e Zika, transmitidas pelo mosquito *Aedes aegypti*, estão entre as doenças mais comuns em ambientes urbanos brasileiros. Seu caráter endêmico resulta em recorrentes impactos na saúde pública. Em Belo Horizonte, o aumento do número de casos notado nos últimos anos preocupa tanto a população quanto as autoridades municipais que, em resposta, intensificaram as ações públicas objetivando sua contenção. Dentre as ações realizadas, as relacionadas ao controle e monitoramento do vetor, tais como vistorias em imóveis, aplicação de inseticidas e mutirões de limpeza, são as que se mostram mais eficientes. Além das citadas, a contabilização de ovos do mosquito depositados em ovitrampas é uma ação de suma importância ao permitir embasar investigação mais refinada da distribuição geográfica dos criadouros dos mosquitos e o direcionamento das demais ações preventivas.

Por volta de 1,8 mil armadilhas localizam-se em pontos estratégicos na malha urbana da cidade, cobrindo um raio de 200 metros cada. Com frequência aproximadamente quinzenal, seu material é coletado e enviado para o Laboratório de Entomologia da Prefeitura de Belo Horizonte (PBH), onde a contagem de ovos é efetuada. (19) Desde o início desse monitoramento, no ano de 2006, estudos são realizados na rica base de dados a disposição da Prefeitura de Belo Horizonte, objetivando descrever a dinâmica do mosquito e realizar previsões relativas aos focos. Entretanto, a alta resolução espaço-temporal das informações pouco foi explorada nos trabalhos realizados.(29)

O objetivo do atual projeto é aliar técnicas de Ciências de Dados e Aprendizado de Máquinas para, em parceria com análises em andamento por parte de pesquisadores da prefeitura, estudar a dinâmica dos criadouros do mosquito e explorar a capacidade preditiva presente nos dados.

Em especial, o emprego de Redes Bayesianas permitirá a aplicação de modelos que consideram e quantificam as incertezas inerentes ao fenômeno analisado. Os resultados serão disponibilizados à Prefeitura de Belo Horizonte, de modo a complementar as análises que a Secretaria Municipal de Saúde de Belo Horizonte (SAMS-BH) realiza.

## 2 Pesquisa Bibliográfica

### 2.1 Área de Estudo

Belo Horizonte, capital do estado de Minas Gerais, estende-se por uma área de 331.354  $km^2$ , dos quais 274,04  $km^2$  são urbanizados. Sua população, de acordo com o censo de 2022 (8), é de 2.315.560 habitantes, resultando em densidade populacional de 6.988,18 pessoas/ $km^2$ . Em relação à concentração de renda, seu índice de Gini é de 0,594, relativamente baixo para os padrões brasileiros (25).

A cidade está situada entre 680 e 1.267 metros acima do nível do mar e seu clima é descrito como tropical subsequente e semi-úmido, caracterizado por pelo menos um mês com temperatura média entre 15 °C e 18 °C e quatro a cinco meses secos ao longo do ano (7). Ela apresenta temperatura média anual de 22,1 °C e precipitação acumulada anual média de 1.578,3 mm (10) com duas estações características, uma estação quente e chuvosa de outubro a março e uma estação seca e mais fria entre de abril e setembro com temperaturas médias e precipitação mensais de, respectivamente 23,4 °C e 231,9 mm, e 20,8 °C e 31,2 mm (10).

### 2.2 Ovitrampas

Ovitrampas, Figura (19), são armadilhas constituídas de um tubo de PVC de aproximadamente 12 centímetros de diâmetro preenchido com infusão de *Panicum maximum* (capim-colonião), responsável pela atração das fêmeas do mosquito *Aedes aegypti*. Imersa na solução, uma placa áspera de coloração escura fixa os ovos dentro do recipiente. As armadilhas são colocadas ao redor de domicílios, em locais sombreados, abrigados da chuva e com menor fluxo de pessoas e animais. Após sete dias de instalação,

as ovitrampas são recolhidas e levadas para análise laboratorial, na qual é feita a classificação dos ovos em três categorias (viável, ressecados e eclodidos) e sua contagem conforme cada categoria. O número de ovos em cada armadilha varia comumente entre 10 e 50, podendo ultrapassar a ordem de milhar em locais de infestação severa.



Figura 1: Exemplo de ovitrampa utilizada em Belo Horizonte - MG (19)

O banco de dados utilizado na pesquisa compreende a contagem de ovos segundo as categorias 'viável', 'ressecados' e 'eclodidos' relativa às coletas de 1825 ovitrampas espalhadas pela malha urbana de Belo Horizonte, conforme Figura 2. As armadilhas estão dispostas em uma malha reticulada com aproximadamente 200 metros de distância entre seus nós, de modo que a distância média entre elas é de aproximadamente 400 metros. O período analisado cobre os anos de 2011 e 2024, não havendo alteração significativa na localização de cada armadilha durante este período.

As armadilhas são instaladas quinzenalmente em um padrão alternado, com instalação em quatro regiões em semanas ímpares e nas regiões restantes em semanas pares. O padrão ocorre ao longo de todo o ano, com exceção de duas semanas no final do ano e durante o Carnaval, que ocorre entre fevereiro e abril. Após coletadas, as armadilhas são levadas para o Laboratório de Entomologia da Prefeitura, onde os ovos são classificados e contabilizados.

### 2.3 Breve Histórico

Em resposta à epidemia de dengue ocorrida no final de 1997 até maio de 1998, o município de Belo Horizonte intensificou o controle do vetor da doença com o início do Programa de Erradicação do Aedes aegypti (PEAa). Como parte dos esforços de contenção às arboviroses, coletas sistemáticas de armadilhas de oviposição, ou ovitrampas, foram introduzidas em 2002. A partir de então, cerca de 1800 armadilhas espalhadas pelo perímetro urbano da cidade em uma grade regular (Figura 2), distando 200 metros entre si, são monitoradas com uma frequência quinzenal. As métricas epidemiológicas obtidas, como o percentual de armadilhas positivas e o número de ovos em cada uma delas, são utilizadas na preparação de relatórios periódicos que auxiliam na elaboração de prognósticos e de propostas de intervenção (20).

Logo nos primeiros anos após a implementação do sistema, análises sobre a dinâmica espacial dos resultados e sobre sua correlação com demais variáveis foram realizadas. O aumento do índice de infestação vetorial nos períodos chuvosos foi confirmado, assim como a sensibilidade das ovitrampas em estações secas, períodos nos quais outros indicadores, como focos larvários, praticamente não são encontrados. (20)

Os resultados de tais análises tanto acrescentam-se aos esforços de trabalhos anteriores de modelar a dinâmica de casos de dengues e prever suas ocorrências (32) quanto embasam novos modelos. Por exemplo, (18) utilizou a média de ovos de julho a outubro de 2009 junto ao Índice de Infestação Predial - porcentagem do número de imóveis positivos dentre os pesquisados - e a dois indicadores de intervenção, proporção de imóveis acessados para controle dos focos e proporção de imóveis não

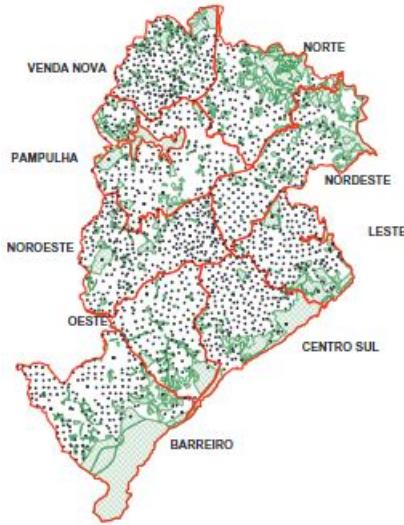


Figura 2: Grade de ovitrampas em Belo Horizonte (19)

acessados por recusa, em um modelo binomial negativo para avaliar a taxa de incidência dos casos notificados de dengue em 2010, por local de residência. Apesar da limitação temporal e espacial do estudo, que foi realizado apenas em três regionais de Belo Horizonte e com dados de ovitrampas no período de seca, ele evidenciou a utilidade dos índices obtidos nas ovitrampas para detecção da presença do vetor, com isso possibilitando seu uso na predição de novos casos.

Estudo estatístico posterior encontrou índices consideráveis de correlação de Pearson entre a média do Número de Ovos nas ovitrampas espalhadas pela cidade e o percentual de ovitrampas positivas ( $r = 0.96, p < 0.01$ ), a Densidade de Ovos em ovitrampas positivas ( $r = 0.96, p < 0.01$ ), a temperatura mensal ( $r = 0.65, p < 0.01$ ) e a precipitação mensal ( $r = 0.54, p < 0.01$ ). Uma Regressão Linear Simples entre a média de ovos nos meses de agosto-setembro e o número de casos anuais no ano posterior foi calculada ( $R = 0.72, p < 0.01$ ). Desse modo, a viabilidade da média de ovos na detecção de flutuações sazonais na população de *Aedes aegypti* e de casos da doença foi apontada, assim como a possibilidade de utilizar dados metereológicos na previsão daquela. (19)

Em um contexto mais amplo, vários estudos têm sido realizados para compreender a dinâmica das arboviroses, variando no que tange ao objetivo do modelo, às técnicas empregadas, às variáveis de entrada e às fontes dos dados. (15) Os modelos para a previsão temporal do número de casos positivos de dengue destacam-se devido à sua popularidade em relação a classificadores de casos positivos e previsores de surtos. Esses modelos, entretanto, não incluem análises sobre a distribuição espacial das doenças, o que limita sua aplicação. Apesar de suprirem esta necessidade, modelos que incluem análise espacial foram pouco explorados. (15)

Nos exemplos existentes de previsão espaço-temporal, a maioria dos estudos considera, além do histórico de casos confirmados da doença, variáveis climáticas como parâmetros do modelo.(31; 1; 33) Porém, parâmetros relacionados a variáveis obtidas por sensoriamento remoto(13), dados de redes sociais (30), consultas em ferramentas de busca (17) e, principalmente, dados relacionados ao monitoramento do vetor (3; 11; 16) também são ocasionalmente utilizados quando disponíveis.

Em relação às técnicas, os modelos de Poisson (6) e modelos de média móvel (ARIMA, SARIMA, ARIMAX) (4; 22; 21) são amplamente utilizados para previsão de séries históricas. Entretanto, constata-se o crescente uso de Redes Neurais Artificiais(23; 12), principalmente Long Short Term Memory (LSTM) (5), Máquinas de Vetores de Suporte (SVM) (28) e modelos baseados em Árvores de Decisão (27).

## 2.4 Estado da Arte

Retornando ao contexto da análise ovitrampas, novos estudos buscam aplicar modelos modernos para aperfeiçoar previsões. Modelos de aprendizado profundo foram utilizados em (26) para prever o Índice de Densidade de Ovos em uma resolução espacial refinada. Para a obtenção do índice, técnicas de suavização espacial e agregação foram aplicadas aos dados das ovitrampas na fase de pré-processamento com intuito de reduzir o efeito da aleatoriedade em pontos individuais e de outliers. O treinamento envolveu o uso de uma janela móvel das 4 semanas anteriores para prever os dados da semana subsequente, sem o uso de variáveis exógenas. Oito modelos foram escolhidos pelo seu amplo uso em previsões epidemiológicas e por sua alta precisão: dois Perceptrons Multicamadas (MLP), três LSTM e três Gated Recurrent Unit (GRU). Dentre eles, um dos modelos LSTM exibiu a melhor generalização. A tentativa da previsão dos valores das ovitrampas sem agregação por estes mesmos modelos apresentou desempenho inferior, visto que, por serem treinados com as médias, tais modelos não conseguiram capturar a dinâmica individual das armadilhas.

Em relação aos dados de ovitrampas de Belo Horizonte, o trabalho mais recente encontrado foi publicado em 2021. Este estudo teve por objetivo avaliar os padrões espaciais e temporais da Incidência de Dengue e do Índice de Positividade de Ovitrampa (OPI), além de analisar a correlação espacial entre essas variáveis. Foram utilizados Global Moran's I e Local Indicator of Spatial Association (LISA) para a identificação de agrupamentos espaciais. Os dados eram relativos ao período de 2007 a 2018 e foram agrupados anualmente e conforme área de abrangência do centro de saúde de cada regional. Como resultado, foram encontrados índices positivos em praticamente todos os anos. Além disso, a distribuição espacial do OPI manteve-se estável ao longo do tempo, um indicativo da presença de criadouros persistentes. Ela contrastava com a distribuição variável da incidência de dengue, insinuando que a baixa presença de ovos não foi um fator limitante para a transmissão da doença. Os próprios autores reconhecem a baixa resolução na escala espacial e a necessidade de considerar outros fatores na análise, como ambiente no qual as armadilhas estão inseridas e fatores socioeconômicos, sugerindo assim novos trabalhos mais refinados nesse sentido. (29)

## 3 Descrição dos Dados

### 3.1 Ovitrampas

Os dados referentes à malha de ovitrampas de Belo Horizonte foram obtidas por meio da submissão de um projeto à prefeitura da cidade, seguindo tutorial disponível em !!!!!!!!!! submetido no dia !!!!!!!!!!.

Após análise do projeto, foram disponibilizados dois grupos de dados. O primeiro, referente aos dados crus coletados diretamente do sistema Prodebel !!!!!!!!!!, no qual a leitura da coleta é digitalizada pelos técnicos da instituição. Esta base continha dados referentes a !!!!!!!!!! armadilhas, com amostra coletadas entre !!!!!!!!!! de 2011 e !!!!!!!!!! de 2024, totalizando !!!!!!!! amostras. Cada linha da base continha informações referentes à uma coleta, como quantidade de ovos eclodidos, secos e !!!!!!!! aferida, data de instalação e de coleta da placa e localização da armadilha !!!!!!!!, somando um total de !!!!!!!!!! colunas.

A segunda base de dados, por sua vez, era resultado do processamento e análises realizadas pela equipe técnica da prefeitura na base descrita anteriormente. Um total de 1339 amostras, consideradas irrecuperáveis devido à ausência do valor de ovos, foram descartadas. Novas colunas referentes à !!!!!!!!!! e categorização das armadilhas foram acrescentadas. Apesar de ainda presentes, os valores faltantes não comprometiam a integridade da amostra e puderam ser tratados em momentos futuros. [Complementar email dilermando]

Além dos dois conjuntos descritos, foram disponibilizados um dicionário com a descrição de cada coluna da primeira base e !!!!!!!!!!, que será disponibilizado no Apêndice !!!!!!!!!!. Demais informações referentes à segunda base, explicações quanto às análises realizadas e esclarecimentos

relativos a valores inconsistentes foram conseguidas por meio de contato direto com representante PBH.

Na tabela 1 encontram-se as colunas do segundo conjunto escolhidas para utilização neste trabalho e sua respectiva descrição. Pontua-se que as colunas de ano, mês e semana referem-se ao ano epidemiológico, que se distingue da divisão anual comum ao ser iniciado no mês de junho, conforme padrão estabelecido pelos técnicos da prefeitura. Esta escolha é feita com intuito de alinhar a divisão anual com o ciclo sazonal do número de ovos, cujo pico encontra-se nos meses de verão. Desse modo, contagens referentes a momentos de alta no mesmo ciclo não são divididas. A convenção de nomenclatura adotada considera o ano inicial do ciclo, seguido pelos dois últimos números do ano seguinte, separados por um underscore. Desse modo, no ano epidemiológico 2016\_17, são consideradas as placas instaladas entre junho de 2016 e julho de 2017. Por sua vez, coluna GerCat refere-se às categorias geradas pela prefeitura para discriminar armadilhas conforme sua contagem histórica de ovos [conferir!!!!!!!]. A cada armadilha é atribuída uma dentre quatro classes de incidência de ovos, B, M A2, A1, respectivamente, baixa, média, alta e muito alta. Detalhes dos cálculos para a atribuição das classes a cada armadilha estão disponíveis no Apêndice !!!!!!!.

Colunas	Descrição
nplaca	Identificador da amostra
novos	Quantidade de ovos coletados na amostra
dtinstal	Data da instalação da armadilha
dtcol	Data da coleta da amostra
narmad	Identificador das armadilhas onde há depósito das placas. Unicidade por local de depósito
anoepid	Ano epidemiológico da amostra. Referente à data de instalação
mesepid	Mês epidemiológico da amostra. Referente à data de instalação
semepi	Semana epidemiológica da amostra. Referente à data de instalação
latitude	Latitude da armadilha
longitude	Longitude da armadilha
GerCat	Categoría da armadilha

Tabela 1: Descrição das colunas utilizadas da base de dados !!!!!!!melhorar

Dentre as colunas escolhidas, nas referentes à localização das armadilhas foram encontrados valores incorretos, em específico coordenadas ausentes ou incoerentes e duas ou mais armadilhas em uma mesma localidade. Após contato com a prefeitura, esclareceu-se que as armadilhas sem valor de latitude e longitude foram desativas e que duas ou armadilhas com mesmas coordenadas estavam instaladas em pontos distintos em parques e [unidades de conservação!!!!!! conferir email dilermando]. Desse modo, amostras com latitude inexistente ou com valores absurdo foram descartadas, armadilhas com mesma localização foram mantidas, porém, um pequeno valor foi adicionado às suas coordenadas para sua distinção na aplicação de métodos de agrupamento. No total, foram descartadas !!!!!!! amostras, restando !!!!!!! para serem trabalhadas, contabilizando !!!!!!! armadilhas distintas, distribuídas conforme mapa da Figura ??.

Dados incorretos foram encontrados também nas colunas de data de instalação e data de coleta das placas, especificamente datas de coleta anteriores à de instalação, datas de coleta posteriores à data de entrega dos dados e datas de instalação e coleta estranhamente distantes. Novamente, o contato com a prefeitura esclareceu que os erros nas datas das amostras tinham provável natureza na inserção dos dados na base e poderiam ser tratadas e corrigidas. Adotou-se tratamento individual para cada amostra incorreta, constando-se prevalência na inserção de datas de coleta. pp sobre o tratamento dos dados problemáticos encontram-se no Apêndice !!!!!!!.

80% entre 14+-1 e 8% entre 28+-1

99% em 7+-1



Figura 3: Mapa com a localização das armadilhas analisadas

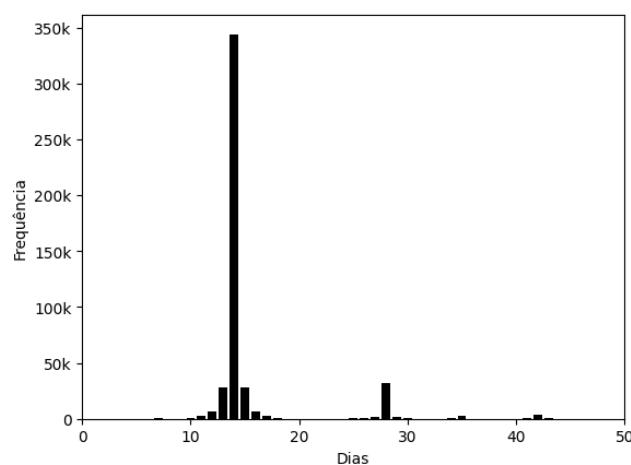


Figura 4: Caption

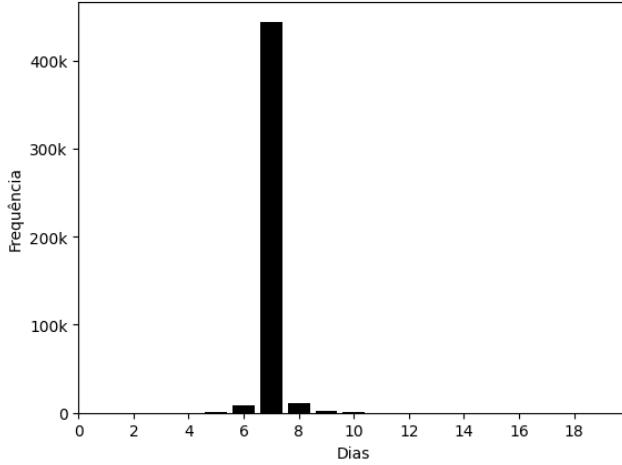


Figura 5: Caption

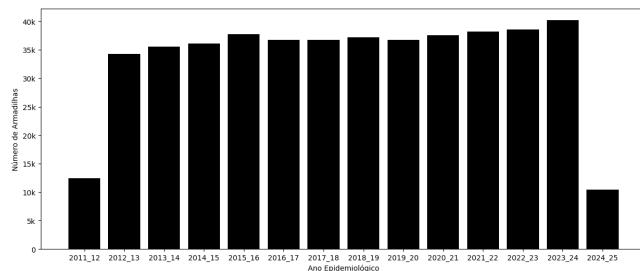


Figura 6: Caption

constante crescente. Números baixos em 2011/12 e 2024/25. Em 2011/12 porque o dataset começou a ser registrado na metade do ano epidemiológico e em 2024/25 porque não havia dados disponíveis no momento da análise

carnaval e natal marcados. Metade dos valores observados

Tabela 2: Category Data with Egg Counts and Means

<b>GerCat</b>	<b>Count</b>	<b>Percentage</b>	<b>Eggs Sum</b>	<b>Eggs Mean</b>
A1	23737	0.050668	2302731.0	97.010195
A2	291379	0.621963	12365097.0	42.436473
B	59910	0.127881	538656.0	8.991087
M	93457	0.199489	1865247.0	19.958344

Número de armadilhas mean 264.363072 min 8.000000 mediana 294.000000 max 311.000000  
 Número de ovos mean 36.440449 min 0.000000 mediana 0.000000 max 4227.000000

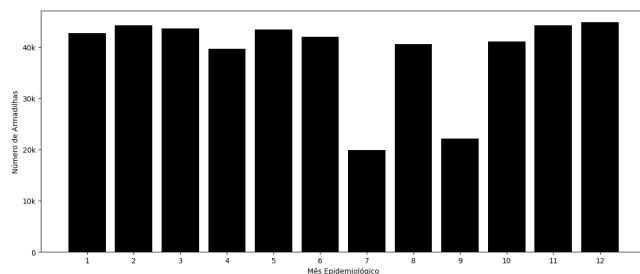


Figura 7: Caption

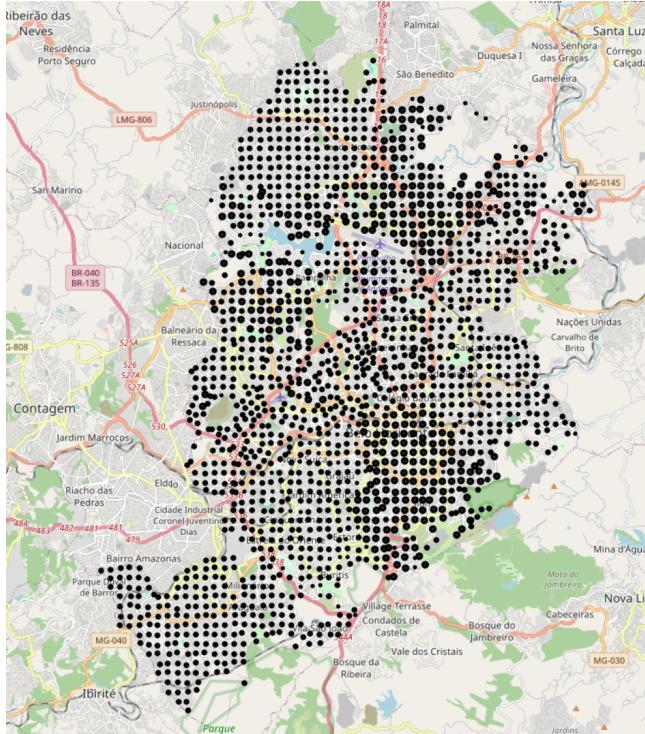


Figura 8: Caption

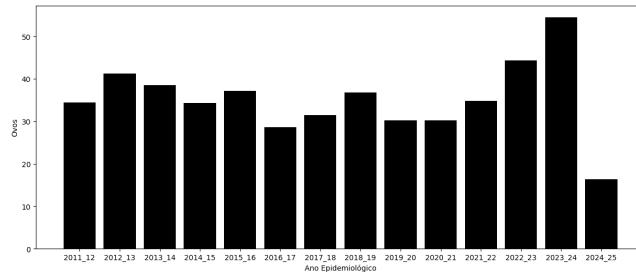


Figura 9: Caption

Número de ovos para armadilhas maiores que 1 mean 75.611233 min 1.0 mediana 48.000000 max 4227.000000

Indício claro dos anos epidemiológicos

Sazonalidade marcada. Não é afetado pelo carnaval e natal

### 3.2 Variáveis exógenas

#### 3.2.1 Dados Metereológicos

Além dos dados das ovitrampas fornecidos pela PBH, serão utilizados dados de temperatura, umidade e pluviometria coletados por estações meteorológicas do Instituto Nacional de Meteorologia (INMET) (9) e por pluviômetros automáticos do Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN) (2), distribuídos conforme mapa da Figura 18. As estações e os pluviômetros cobrem a região metropolitana de Belo Horizonte em uma malha não regular, coletando dados por períodos com início distinto, muitas vezes não cobrindo o período de coleta das ovitrampas.

Inversamente proporcional à média de ovos

Tendo em vista a incompatibilidade espacial entre esta malha e a malha das armadilhas e a inexistência de amostras meteorológicas de parte das estações para todo o período de coleta de ovitrampas, será necessário processamento prévio para conformação dos dados meteorológicos aos entomológi-

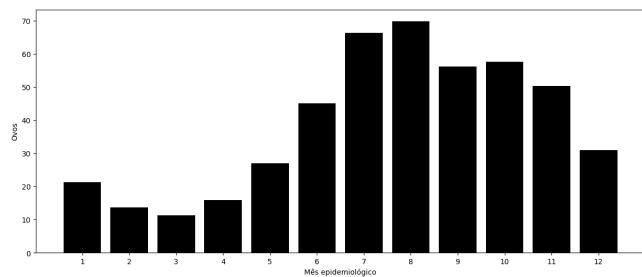


Figura 10: Caption

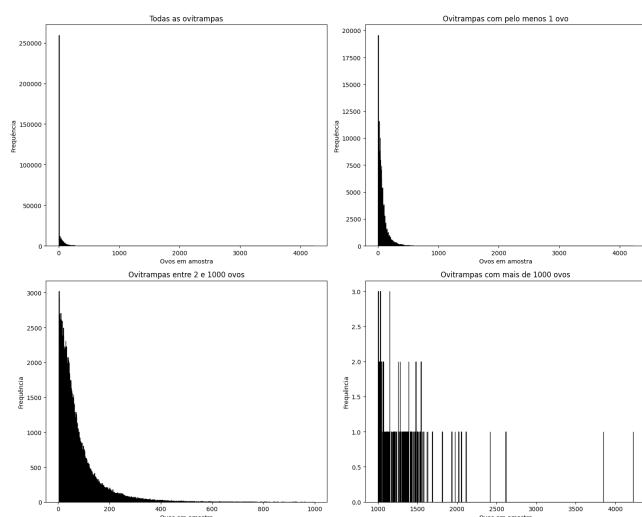


Figura 11: Caption

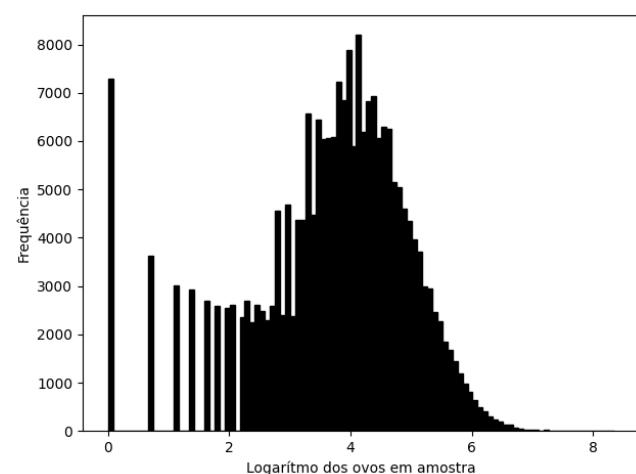


Figura 12: Caption

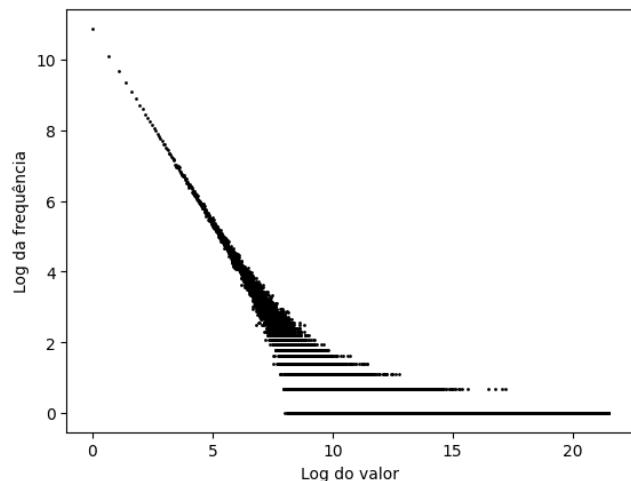
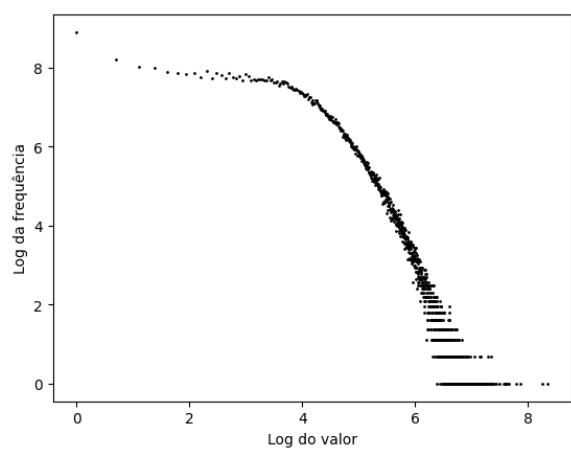
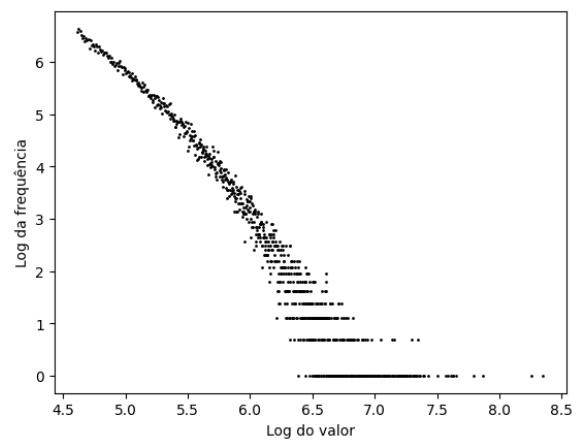


Figura 13: Caption



(a) First figure description.



(b) Second figure description.

Figura 14: Overall caption for the figures.

cos. Para tal, propõe-se associar valores relativos à temperatura e à precipitação a cada amostra da contagem de ovos em função da distância da ovitrampa às estações meteorológicas disponíveis no momento da coleta.

Além disso, os dados dos pontos de coleta foram registrados com taxas amostrais intradiárias variáveis e, no caso de Estações Convencionais, com frequência irregular (diariamente, às 00h, 12h e 18h). Com o intuito de padronizar as informações em taxas iguais, será realizado o agrupamento dos dados meteorológicos por média diária. Por fim, para a adequação das taxas de coleta dos dados entomológicos e meteorológicos ocorrerá paralelamente por meio de duas heurísticas distintas, a agregação das variáveis meteorológicas em amostras quinzenais e a interpolação do número de ovos por ovitrampas para taxas diárias. Ambas serão utilizadas nos diferentes modelos e comparadas conforme as métricas de qualidade descritas em frente.

## 4 Metodologia

Xcor da série de diferenças normalizadas de cada sinal.

Em 1, 420K Em 2, 373K Em 8, amostras caem pela metade 213K Em 14, o número de amostras cai para 95K

i - Número máximo de armadilhas vizinhas escolhidas para matriz. Define range de 0 a i, sendo

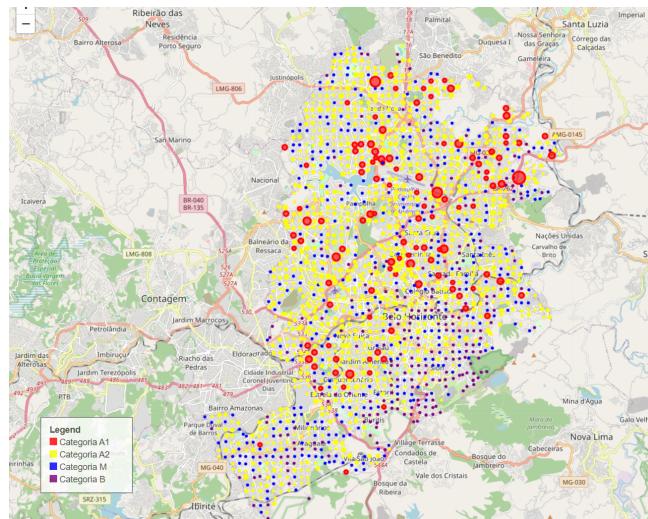


Figura 15: Caption

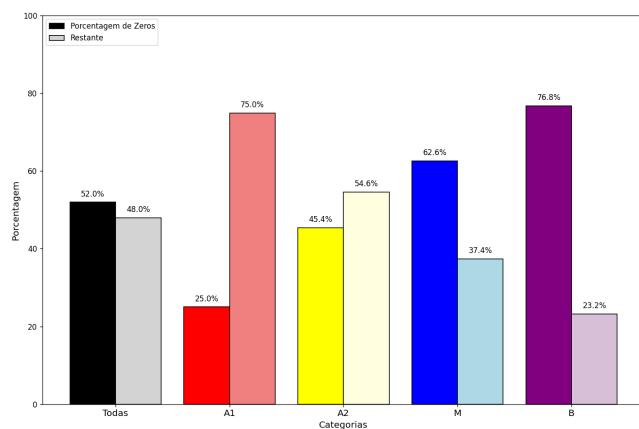


Figura 16: Caption

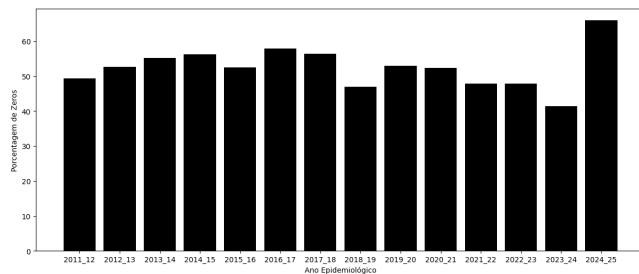


Figura 17: Caption



Figura 18: Localização das estações meteorológicas e pluviômetros na região metropolitana de Belo Horizonte (9)

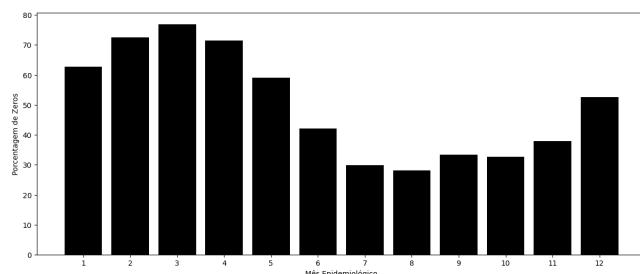


Figura 19: Caption

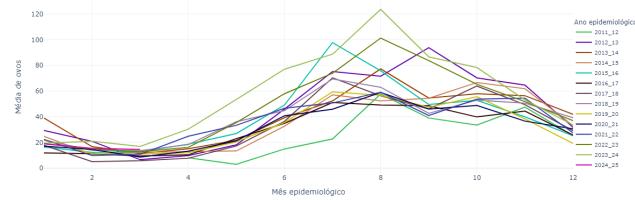


Figura 20: Caption

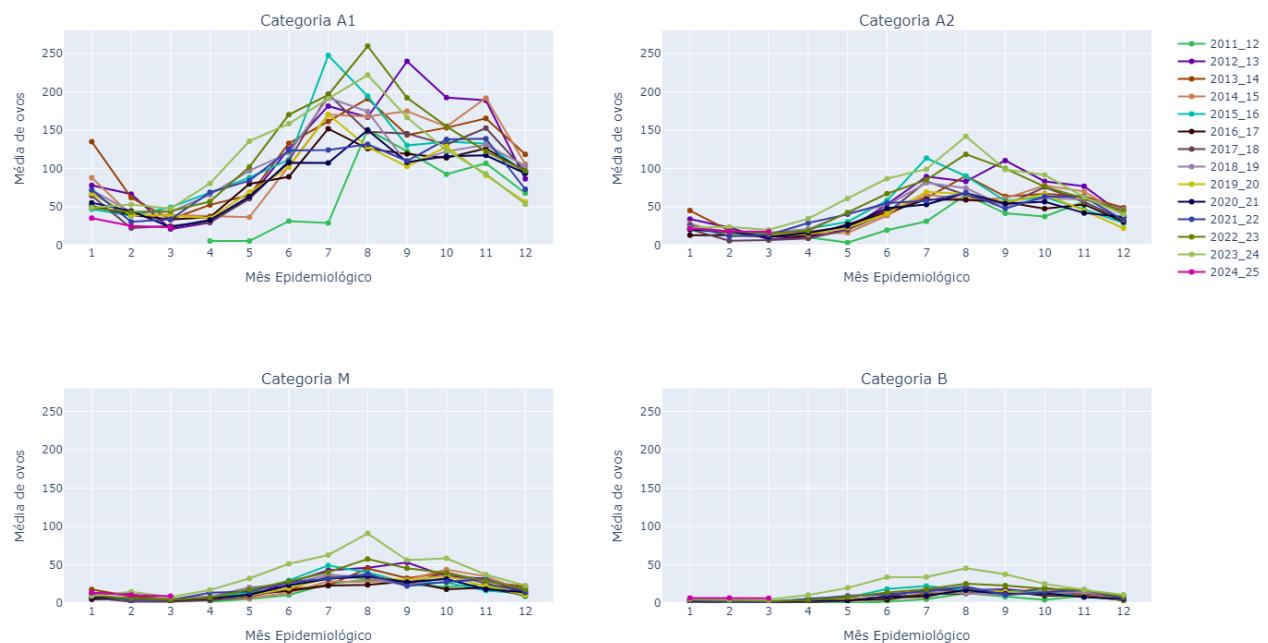


Figura 21: Caption

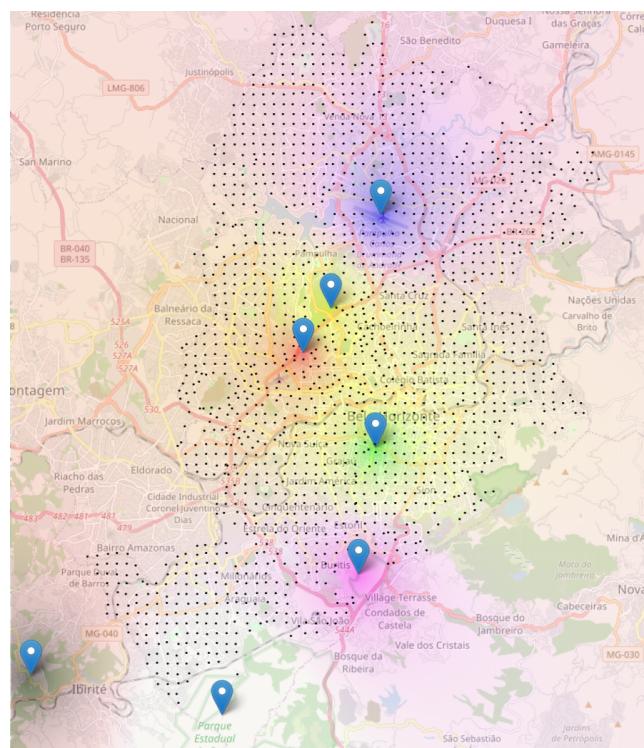


Figura 22: Caption

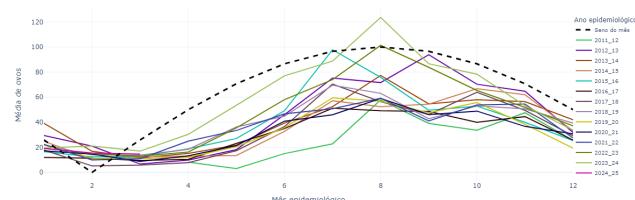


Figura 23: Caption

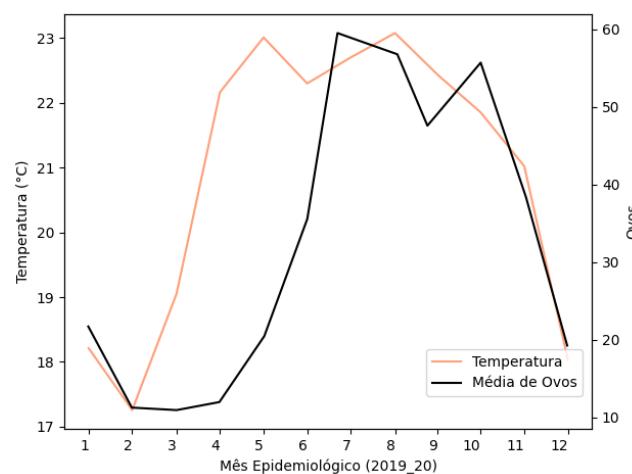


Figura 24: Caption

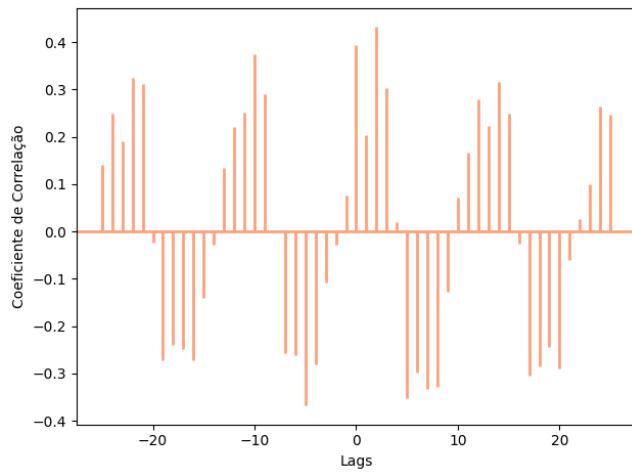


Figura 25: Caption

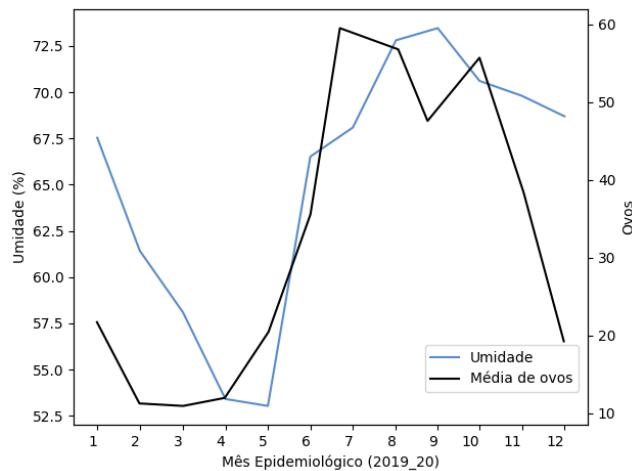


Figura 26: Caption

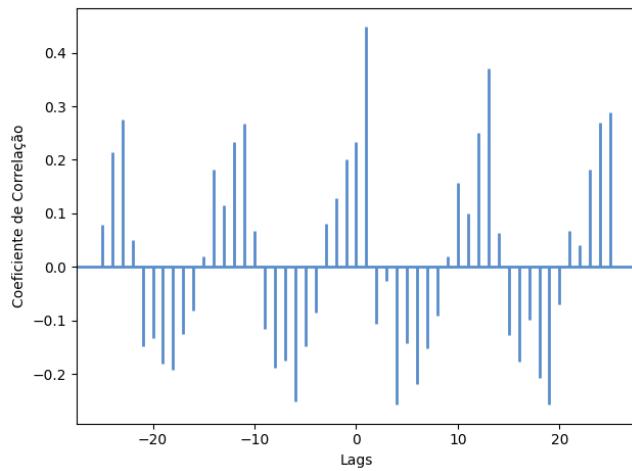


Figura 27: Caption

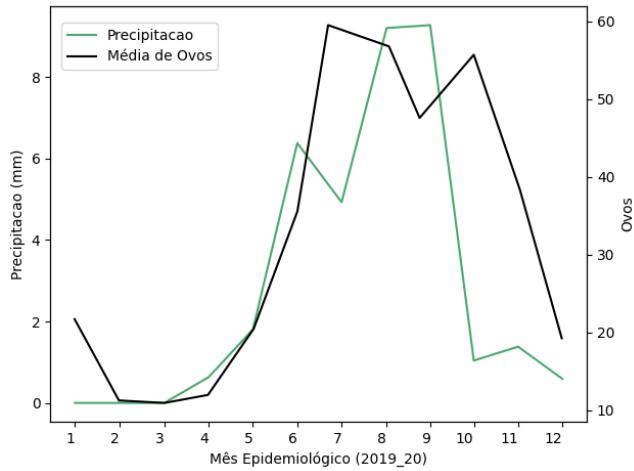


Figura 28: Caption

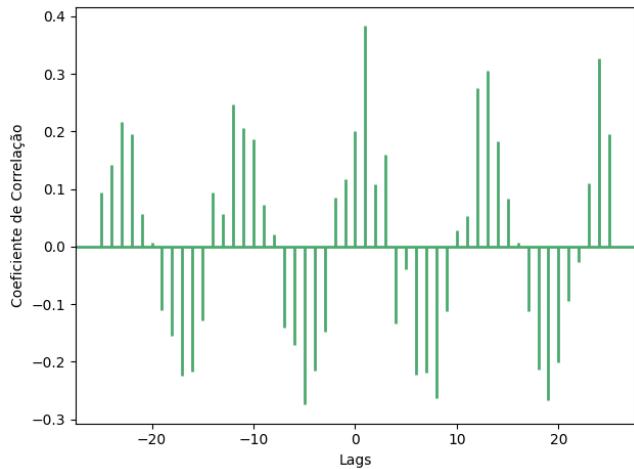


Figura 29: Caption

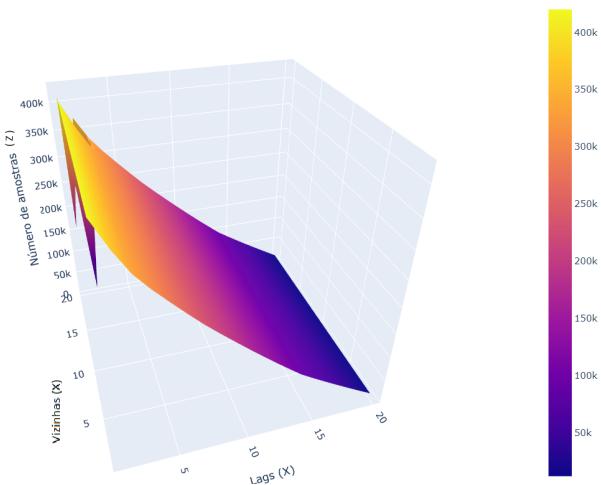


Figura 30: Caption

Column Name	Description
nplaca	Identificador da amostra
novos	Quantidade de ovos coletados na amostra
trap{i}_lag{j}	Quantidade de ovos coletados na armadilha vizinha i com um lag de j
latitude{i}	Latitude da armadilha vizinha i
longitude{i}	Longitude da armadilha vizinha i
days{i}_lag{j}	Diferença entre a data de coleta da amostra e da amostra vizinha i com um lag de j
mesepid	Mês epidemiológico de instalação da amostra. Transformado em variável categórica
sin_mesepi	Seno do mês epidemiológico de instalação da amostra
semeipi	Semana epidemiológica de instalação da amostra
semeipi2	Quadrado da semana epidemiológica de instalação da amostra
sin_semeipi	Seno da semana epidemiológica de instalação da amostra
anoepid	Ano epidemiológico de instalação da amostra
temp_expo	Tempo de exposição da amostra (data de coleta - data de instalação)
zero_perc	Porcentagem de amostras com zero ovos para aquela armadilha desde o seu uso
Temperatura_previsao	Temperatura da armadilha nas datas de exposição
Precipitacao_previsao	Precipitação da armadilha nas datas de exposição
Umidade_previsao	Umidade da trilha nas datas de exposição
Temperatura_week_bfr_mean	Temperatura média da semana anterior à instalação da amostra
Precipitacao_week_bfr_mean	Precipitação média da semana anterior à instalação da amostra
Umidade_week_bfr_mean	Umidade média da semana anterior à instalação da amostra

Tabela 3: Tabela com os dados das amostras e armadilhas

0 a própria armadilha, 1 a primeira mais próxima e i a i-ésima armadilha mais próxima. j - Número máximo de lags escolhidos para matriz. Define range de 1, a amostra imediatamente anterior, até a j-ésima amostra. À medida que estes valores são aumentados, principalmente o número máximo de lags, o número de amostras disponível é reduzido devido ao descarte dos NANs.

Truncado em 1000, com jitter de 5

	Anterior 0	Anterior 1
Atual 0	0.375415	0.161822
Atual 1	0.159960	0.302803

Tabela 4: Matriz de Confusão

Dataset lags 1 e Vizinhos 1: máximo de amostras possíveis. Prevalência entre na manutenção da ausência, mas com manutenção da presença também comum. 67.5% de acerto com todas as amostras

#### 4.1 Seleção de entradas

A seleção de entradas para os modelos consistirá em deslocar a Série Temporal de contagem de ovos de uma armadilha por diferentes atrasos (ou *lags*, em inglês) e selecionar os  $T_0$  melhores atrasos conforme o critério de Correlação de Pearson. Desse modo, para a prever da contagem  $y_t$ , no tempo t, os modelos receberão como entrada os valores  $y_{t-lag[i]}$ , sendo  $lag[i]$  o i-ésimo termo do vetor de atrasos, de tamanho  $T_0$ . Este mesmo método será aplicado para as  $E_{ST}$  variáveis exógenas representadas por séries temporais nos  $T_1$  atrasos.

De modo análogo, o I de Moran, uma adaptação da correlação de Pearson para dados espaciais (24), será utilizado como medida da correlação da ovitrampa analisada com os K vizinhos mais próximos.

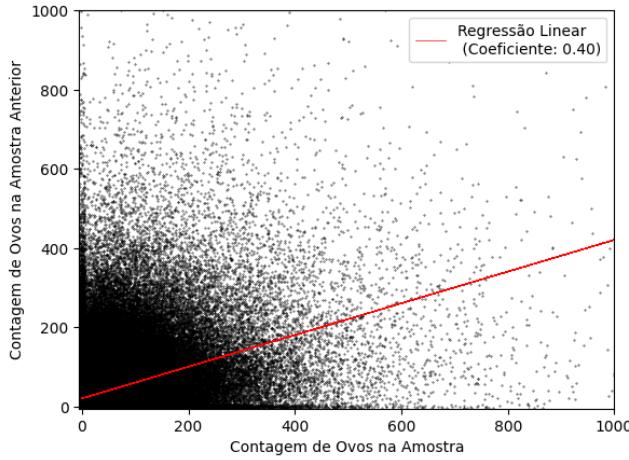


Figura 31: Caption

próximos, selecionando  $T_2$  atrasos. Ao final,  $T_0 + E_{ST} \cdot T_1 + K \cdot T_2 + E_C$  entradas serão utilizadas nos modelos descritos abaixo, sendo  $E_C$  as variáveis exógenas categóricas.

## 4.2 Modelos de Aprendizado de Máquina

Neste estudo, exploraremos dois grupos distintos de modelos de Aprendizado de Máquina para previsão da contagem de ovitrampas. O primeiro grupo consiste em diferentes modelos Multilayer Perceptron (MLP), um tipo de Rede Neural Artificial Feedforward tradicional composta por múltiplas camadas de neurônios interconectadas, Figura ???. A camada de entrada recebe as variáveis escolhidas, enquanto as camadas intermediárias, ou camadas ocultas, permitem ao MLP aprender representações complexas dos dados de entrada. Finalmente, a camada de saída gera a resposta final modelo após aplicar uma função de ativação. O número de camadas ocultas e neurônios serão alterados para criar diferentes modelos a serem testados.

O segundo grupo comprehende modelos Long Short-Term Memory (LSTM), Figura ???, uma classe de Redes Neurais Recorrentes (RNN) que se tornou popular na literatura de modelagem da dengue devido ao seu bom desempenho e alta praticidade (14). Este é um tipo de modelo que utiliza sua própria saída no tempo  $t$  como entrada para a previsão do tempo  $t+1$  e é capaz de identificar automaticamente as tendências de longo prazo e flutuações de curto prazo de séries temporais. Ele é projetado especificamente para evitar os problemas de desaparecimento e explosão de gradiente em sequências temporais longas e possui estruturas especializadas que permitem armazenar, recuperar e esquecer informações ao longo do tempo. Assim como no caso do MLP, o número de camadas ocultas e de neurônios diferenciará os modelos neste grupo.

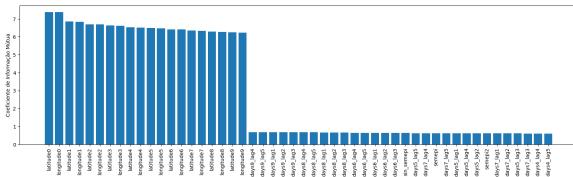
## 4.3 Métricas

A fim de quantificar a variância entre os valores reais da série de ovos e os valores previstos, o desempenho dos modelos será avaliado usando a Raiz do Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE), duas métricas amplamente utilizadas na literatura. (14) (26)

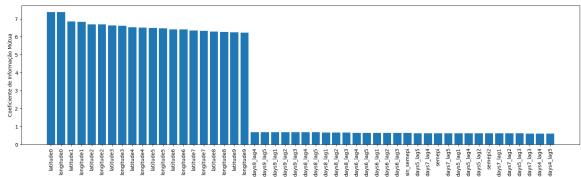
Dois modelos serão utilizados como referência para avaliar o desempenho dos demais. Um modelo ingênuo, que repete o valor anterior da armadilha como predição, e um modelo SARIMAX, comumente utilizado para esse fim (14). SARIMAX é um modelo estatístico utilizado em previsões de séries temporais estacionárias, que inclui variável autorregressiva (AR), média móvel (MA), integração (I), variáveis externas (X) e um componente sazonal (S) do histórico da série temporal.

## 4.4 Seleção de Features

lags = 14, vizinhos = 20 normalizado treinamento lat and long



(a) First figure description.



(b) Second figure description.

Figura 32: Overall caption for the figures.

Curiosamente, lat e long são muito correlacionados

[‘mesepid\_4.0’, ‘mesepid\_7.0’, ‘mesepid\_6.0’, ‘mesepid\_5.0’, ‘mesepid\_3.0’, ‘mesepid\_2.0’] baixo valor para lags além de 5, sem uma estrutura temporal relevante para as vizinhas, valor foi baixo a partir do lag 2

lags = 5, vizinhos = 10 ['mesepid\_1.0' 'mesepid\_2.0', 'mesepid\_3.0', 'mesepid\_4.0', 'mesepid\_5.0', 'mesepid\_6.0', 'mesepid\_7.0', 'mesepid\_10.0', 'mesepid\_11.0', 'mesepid\_12.0'] - janeiro e fevereiro

valor dos novos filter > 0.1

## Dados meteorológicos com valor alto

mesepid

classificação bool input

outra tentativa: stepwise no modelo logístico. Pouca diferença de acurácia

truncando em 100 e normalizando

Resultados dos modelos: Initial: Classification 'logistic' s exeterno 70.0'logistic' c exeterno - 70.3Regression 'linear' s exeterno 30.2 (27.3) 'linear' c exeterno 30.1(27.3) - semana anterior 'linear' c exeterno 30.1(27.1)

Regression (truncando em 100): 'linear' 30.1(27.3) 'Naive<sub>r</sub>eg'38.7(35.4)'svr'42.3(36.9) – sem variações nos parâmetros 'random<sub>f</sub>orest<sub>r</sub>eg'30.2(25.4)'catboost<sub>r</sub>eg'30.0(26.9)'mlp<sub>r</sub>eg'(regression)48.0(

Classification: Porcentagem de zeros - 54.0% 'Naive' - 66.3'logistic' - 70.3'mlp'  
 $(10,10,5)$  - 66.1(71.6) (10,10,5) - 70.4(72.7) (50, 25, 25, 5) - 70.1 (72.6)  
 'randomforest'70.2(100)di ferenamaxima0.1'svm'59.3(64.6)di ferenamaxima14

*melhor* 70.0'catboost'70.8(73.3) diferenamaxima0.33<sub>clases</sub>

*perc<sub>zero</sub>'logistic3c'72.5(75.6)'Naive3c'65.2(68.1)'mlp3c'72.6(76.0)30epochassemudanca, de10<sup>-5</sup>)nloss, adam, relu,' random\_forest3c'72.5(78.2)'svm3c'71.6(74.7)'catboost3c'72.7(76.0)'linear3c'59*

Matriz de confusão [[27759 6580] [11545 16328]]

Sensibilidade (Acertar 1) 0.5857998780181538 Especificidade (Acertar 0) 0.808381140976732

0.07  
0.06  
0.05  
0.04  
0.03  
0.02  
0.01

Figure 1. The distribution of the number of nodes in the network. The x-axis shows the number of nodes, and the y-axis shows the probability density.

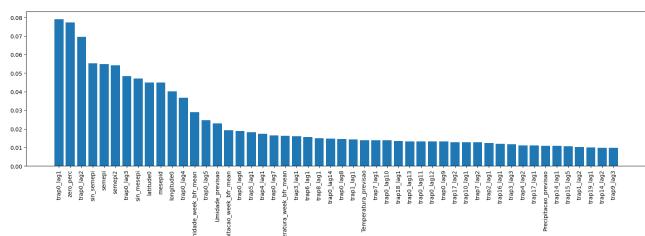


Figura 33: Caption

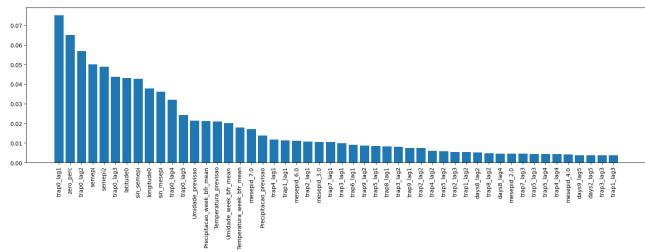


Figura 34: Caption

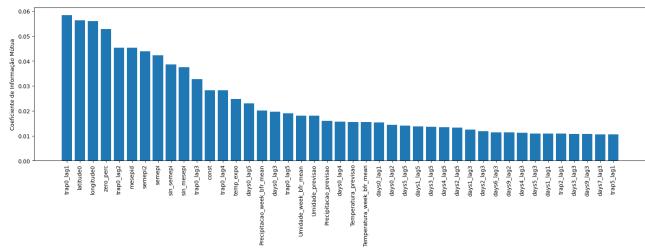


Figura 35: Caption

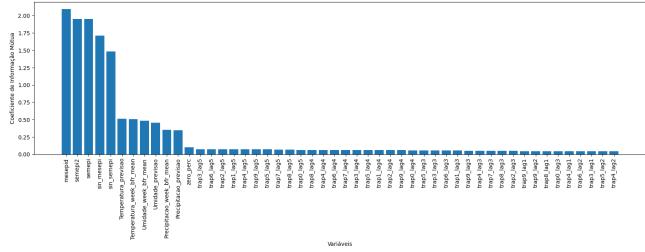


Figura 36: Caption

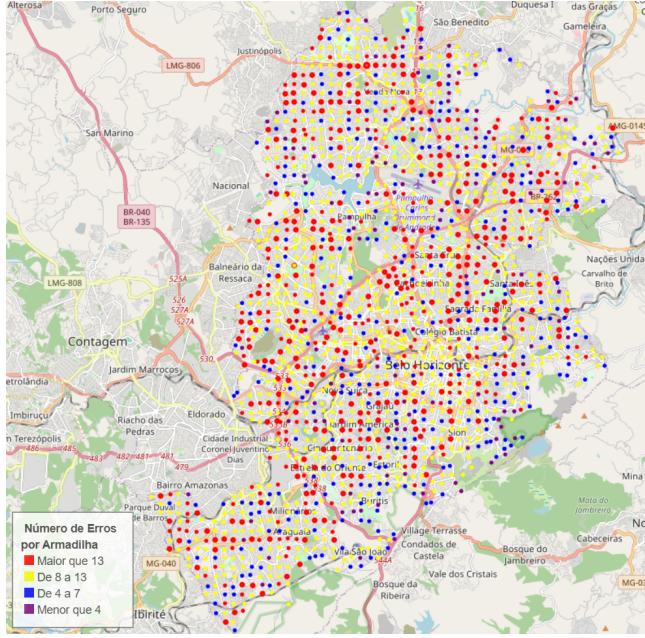


Figura 37: Caption

## 4.5 Resíduos

### 4.5.1 Classificação de Presença

Catboost para classificação: acc 70.9%

Limites da legenda do mapa definidos para dividir quartis

Erros mês	1.0	0.293301	2.0	0.292192	3.0	0.257242	4.0	0.293707	5.0	0.341650	6.0	0.277637	7.0
0.210542	11.0	0.317979	12.0	0.325924									
ano	2022_2	30.277261	2023_2	40.303385	2024_2	50.294327							
Cat A1	0.267086	A2	0.309356	B	0.224403	M	0.284827						

### 4.5.2 Regressão

Limites da legenda do mapa definidos para dividir quartis

Erros clas:	mês	1.0	18.828738	2.0	16.965916	3.0	16.083753	4.0	19.731593	5.0	27.019493	6.0
30.031129	7.0	30.348795	11.0	29.779346	12.0	24.254281						
ano	2022_2	320.5439522023_2	424.8956272024_2	517.968624								
Cat A1	29.684952	A2	24.904288	M	17.537407	B	11.719497					

Comportamento típico de modelo com forte influência da variável autoregressiva. Nota-se dificuldade de prever picos ou armadilhas constantes em 0. Outro comportamento curioso é um aparente limite máximo para valores preditos de acordo com a Classe. Para Classes B e M, os sináis analisados raramente ultrapassavam a faixa de 30 ovos, apesar do valor real ter ultrapassado esta faixa diversas vezes, podendo inclusive alcançar o limite máximo de 100 ovos. Por outro lado, as previsões das Classes A2 e A1 acompanham com maior precisão picos altos, embora acompanhem com menor precisão valores baixos do sinal real.

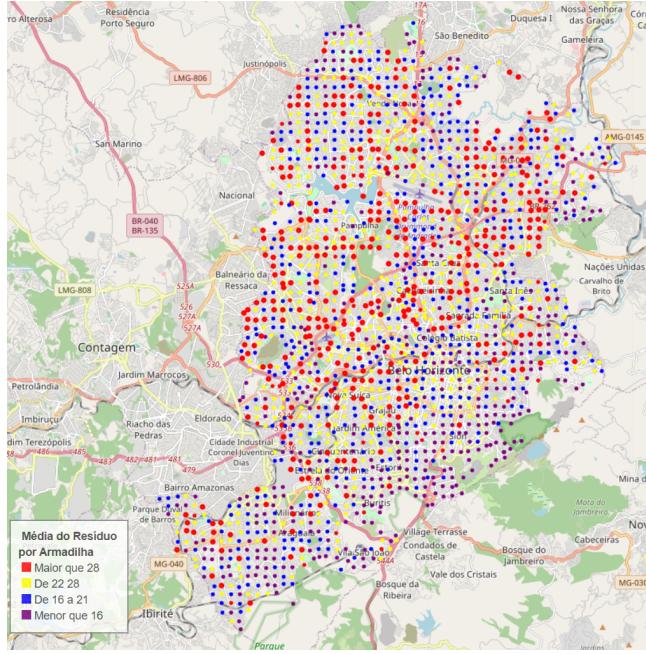


Figura 38: Caption

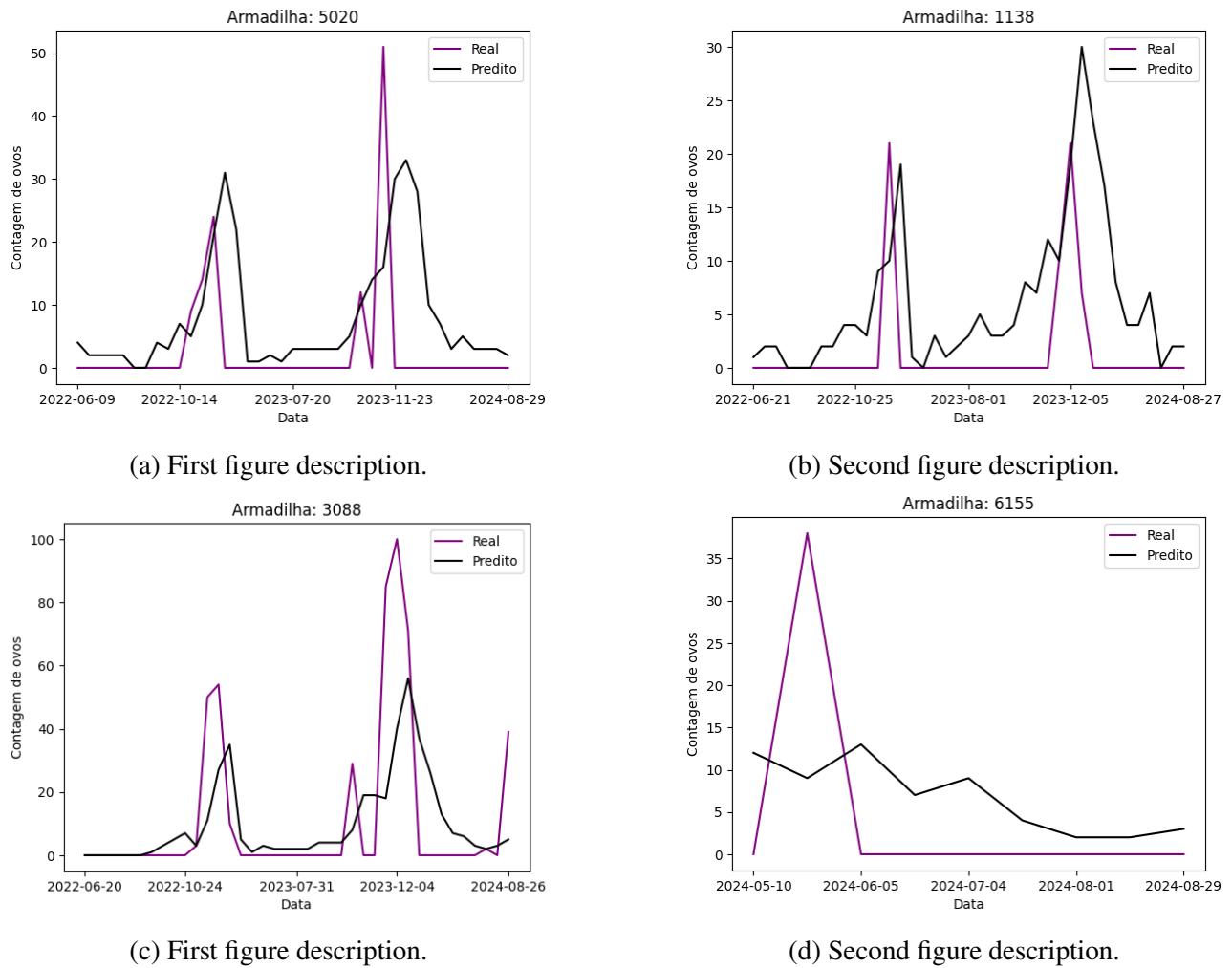
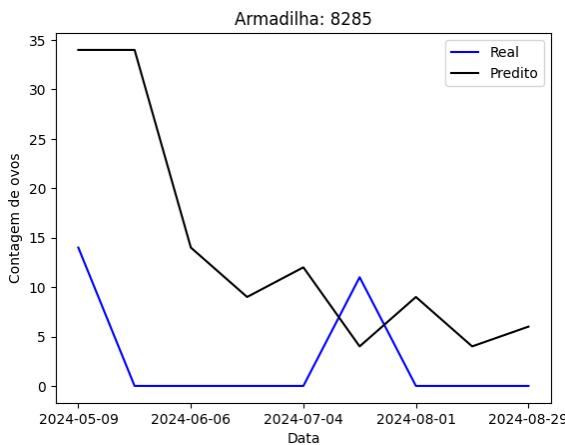
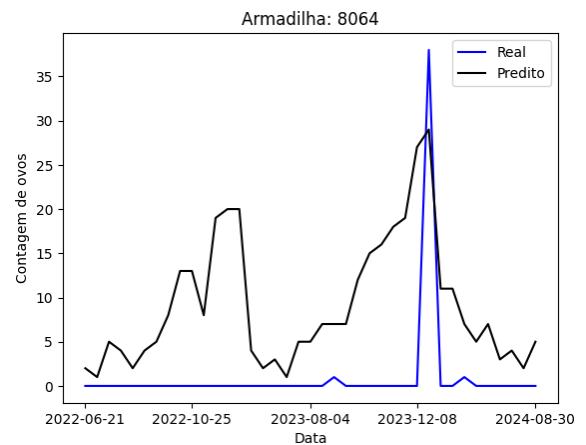


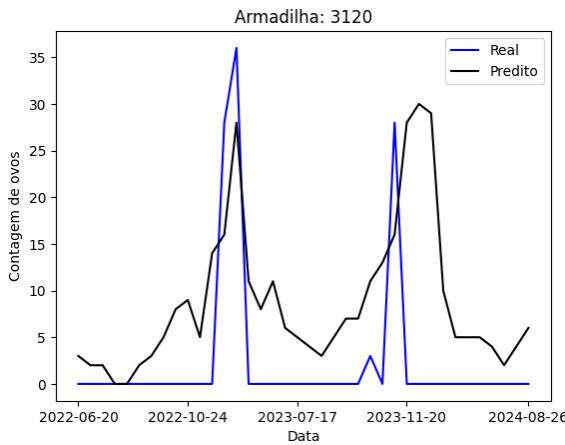
Figura 39: Overall caption for the figures.



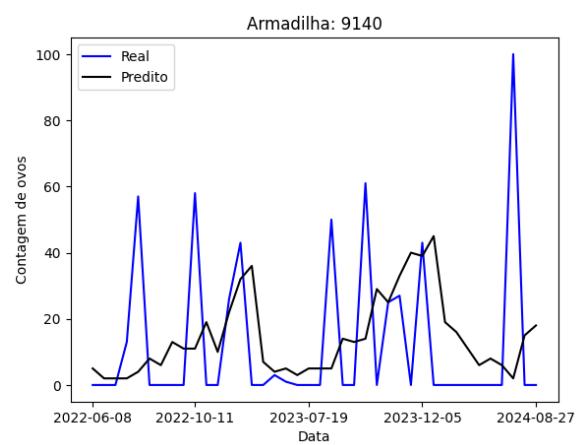
(a) First figure description.



(b) Second figure description.



(c) First figure description.



(d) Second figure description.

Figura 40: Overall caption for the figures.

#### 4.5.3 Cross Validation

## 5 Conclusão

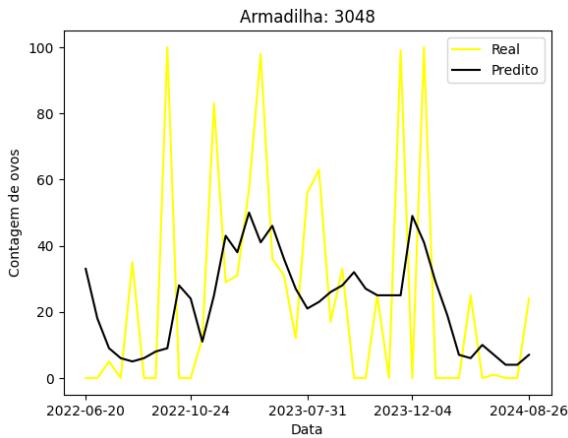
## 6 Apêndice

### 6.1 Tratamento de Valores Inconsistentes

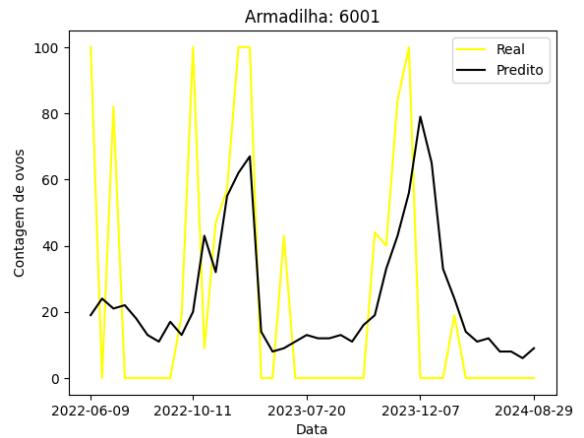
valores extremos foram tratados individualmente e, caso não possível, incorporados ao dataset como parte das imprecisões referentes ao processo de coleta, dado que representam porcentagem ínfima do dataset

### 6.2 Modelos iniciais

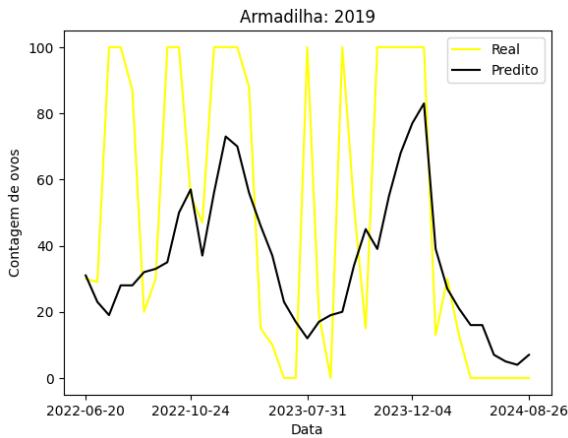
1: lags, days, lat long 68.9 2: semepid 68.9 (sem estrutura) 3: perc<sub>zero</sub>69.05 : semedpid, semepid2, sin<sub>semepid</sub>, (lat, long)daarmadilha70.06 : 100truncado70.01 : step70.0'logistic'sexterno70.0'logistic'cexterno - 70.3



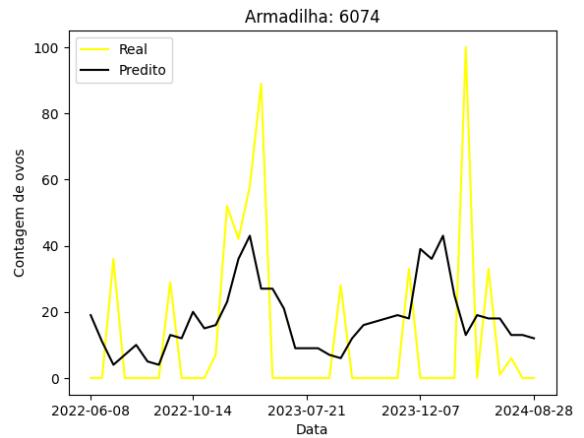
(a) First figure description.



(b) Second figure description.



(c) First figure description.



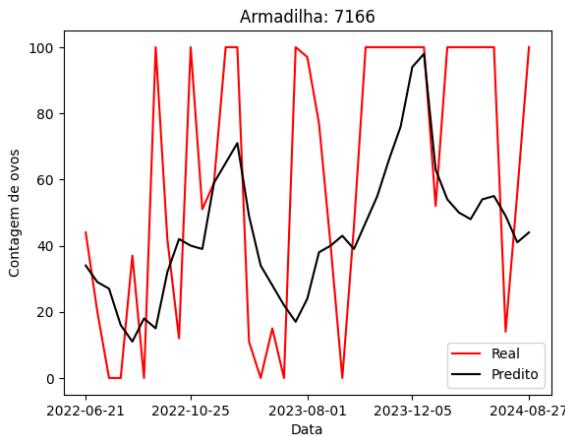
(d) Second figure description.

Figura 41: Overall caption for the figures.

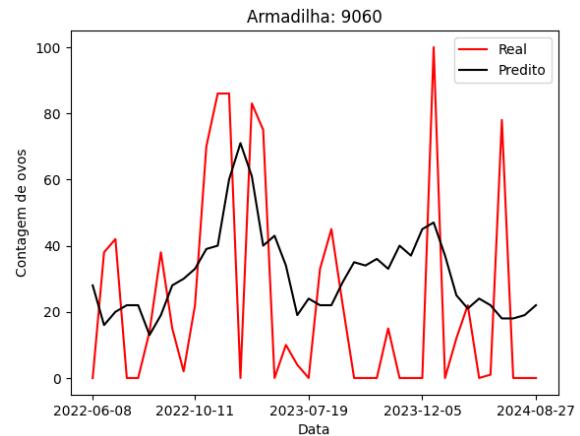
## 7 Bibliografia

### Referências

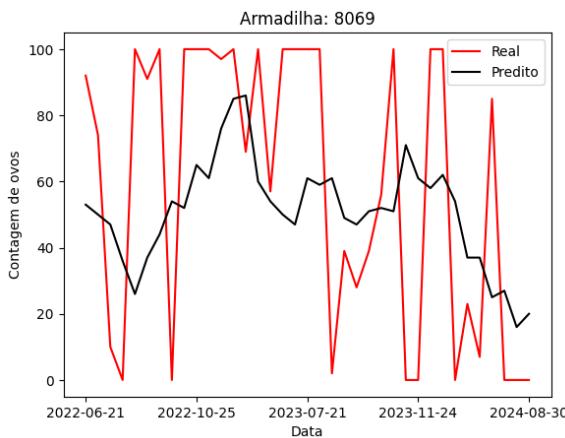
- [1] Thaddeus M Carvajal, Katherine M Viacrusis, Lara Fides T Hernandez, Howell T Ho, Divina M Amalin, and Kozo Watanabe. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan manila, philippines. *BMC infectious diseases*, 18:1–15, 2018.
- [2] Centro Nacional de Monitoramento e Alertas de Desastres Naturais. Mapa interativo de monitoramento de desastres naturais. <https://mapainterativo.cemaden.gov.br/#>, 2024. Acesso em: 10 jun. 2024.
- [3] Fong-Shue Chang, Yao-Ting Tseng, Pi-Shan Hsu, Chaur-Dong Chen, Ie-Bin Lian, and Day-Yu Chao. Re-assess vector indices threshold as an early warning tool for predicting dengue epidemic in a dengue non-endemic country. *PLoS neglected tropical diseases*, 9(9):e0004043, 2015.
- [4] Romrawin Chumpu, Nirattaya Khamsemanan, and Cholwich Nattee. The association between dengue incidences and provincial-level weather variables in thailand from 2001 to 2014. *Plos one*, 14(12):e0226945, 2019.
- [5] Anjelus Ronald Doni and Thankappan Sasipraba. Lstm-rnn based approach for prediction of dengue cases in india. *Ingénierie des Systèmes d'Information*, 25(3), 2020.



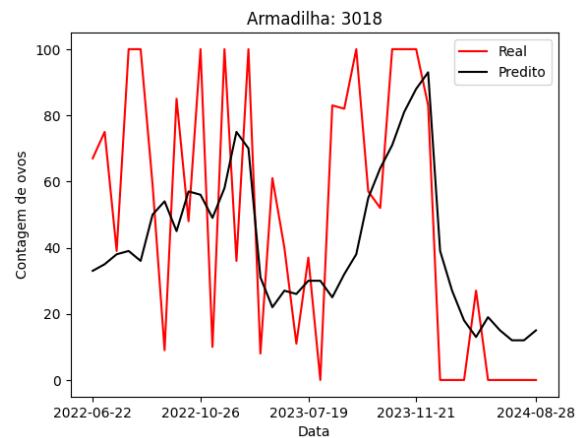
(a) First figure description.



(b) Second figure description.



(c) First figure description.



(d) Second figure description.

Figura 42: Overall caption for the figures.

- [6] Subrata Ghosh, Santanu Dinda, Nilanjana Das Chatterjee, Kousik Das, and Riya Mahata. The spatial clustering of dengue disease and risk susceptibility mapping: an approach towards sustainable health management in kharagpur city, india. *Spatial Information Research*, 27(2):187–204, 2019.
- [7] Instituto Brasileiro de Geografia e Estatística. Mapa climático do brasil - 2002. [https://geftp.ibge.gov.br/informacoes\\_ambientais/climatologia/mapas/brasil/Map\\_BR\\_clima\\_2002.pdf](https://geftp.ibge.gov.br/informacoes_ambientais/climatologia/mapas/brasil/Map_BR_clima_2002.pdf), 2002. Acesso em: 08 jun. 2024.
- [8] Instituto Brasileiro de Geografia e Estatística. Panorama de Belo Horizonte, MG. <https://cidades.ibge.gov.br/brasil/mg/belo-horizonte/panorama>, 2023. Acesso em: 07 jun. 2024.
- [9] Instituto Nacional de Meteorologia. Mapas meteorológicos do brasil. <https://mapas.inmet.gov.br/>. Acesso em: 10 jun. 2024.
- [10] Instituto Nacional de Meteorologia. Normais climatológicas do brasil. <https://portal.inmet.gov.br/normais>. Acesso em: 10 jun. 2024.
- [11] QL Jing, Q Cheng, JM Marshall, WB Hu, ZC Yang, and JH Lu. Imported cases and minimum temperature drive dengue transmission in guangzhou, china: evidence from arimax model. *Epidemiology & Infection*, 146(10):1226–1235, 2018.
- [12] Benjapuk Jongmuenwai, Sudajai Lowanichchai, and Saisunee Jabjone. Comparision using data mining algorithm techniques for predicting of dengue fever data in northeastern of thailand. In

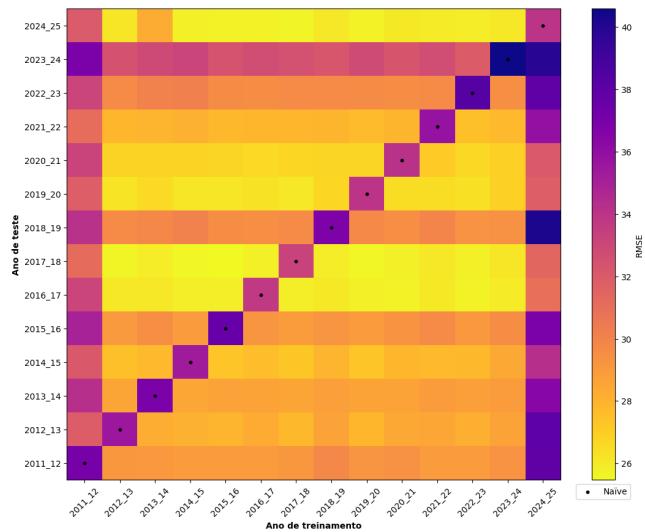


Figura 43: Caption

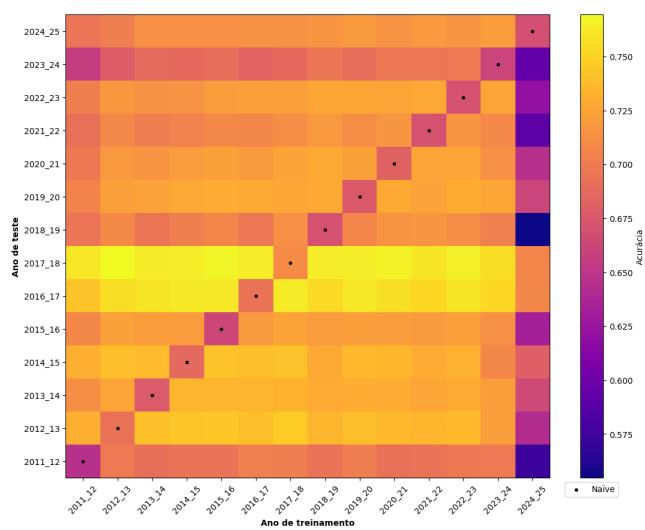


Figura 44: Caption

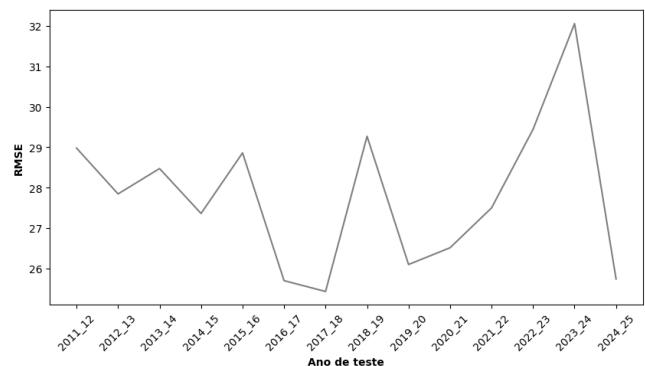


Figura 45: Caption

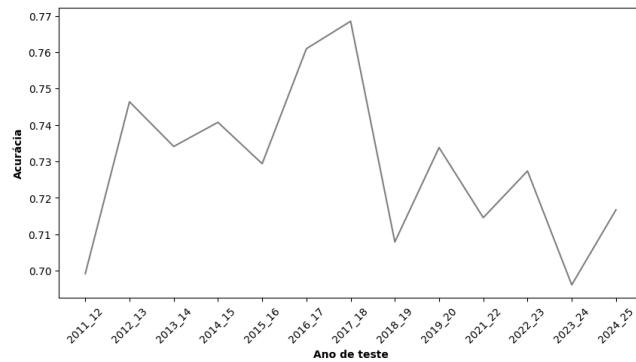


Figura 46: Caption

Tabela 5: sample rate and Corresponding Values

Days	Value
62	1.0
63	5.0
65	1.0
69	2.0
70	35.0
71	3.0
73	1.0
74	1.0
77	23.0
83	1.0
84	8.0
87	1.0
91	2.0
98	3.0
100	1.0
112	1.0
114	1.0
126	11.0
137	1.0
153	1.0
154	2.0
181	1.0
196	8.0
210	4.0
224	1.0
245	3.0
294	1.0
503	1.0
735	1.0

2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pages 532–535. IEEE, 2018.

- [13] N Kerdprasop, K Kerdprasop, and P Chuaybamroong. Computational intelligence and statistical learning performances on predicting dengue incidence using remote sensing data. *Adv Sci*

Tabela 6: dtexpo

Days	Frequency
51	2
56	2
61	2
62	2
63	1
69	2
79	1
83	2
89	1
114	1
119	2
123	2
133	1
136	2
137	1
149	1
296	1
371	1
415	2
3296	1

*Technol Eng Syst J*, 5:344–50, 2020.

- [14] Zhichao Li and Jinwei Dong. Big geospatial data and data-driven methods for urban dengue risk forecasting: a review. *Remote Sensing*, 14(19):5052, 2022.
- [15] Clárisse Lins de Lima, Ana Clara Gomes da Silva, Giselle Machado Magalhães Moreno, Cecilia Cordeiro da Silva, Anwar Musah, Aisha Aldosery, Livia Dutra, Tercio Ambrizzi, Iuri VG Borges, Merve Tunali, et al. Temporal and spatiotemporal arboviruses forecasting by machine learning: a systematic review. *Frontiers in Public Health*, 10:900077, 2022.
- [16] K-K Liu, T Wang, X-D Huang, G-L Wang, Yao Xia, Y-T Zhang, Q-L Jing, J-W Huang, X-X Liu, J-H Lu, et al. Risk assessment of dengue fever in zhongshan, china: a time-series regression tree analysis. *Epidemiology & Infection*, 145(3):451–461, 2017.
- [17] S Morsy, TN Dang, MG Kamel, AH Zayan, OM Makram, M Elhady, K Hirayama, and NT Huy. Prediction of zika-confirmed cases in brazil and colombia using google trends. *Epidemiology & Infection*, 146(13):1625–1627, 2018.
- [18] Elsa Maria Nphantumbo, José Eduardo Marques Pessanha, and Fernando Augusto Proietti. Title of the article. *Revista Médica de Minas Gerais*, 22(3):265–273, Jul/Set 2012.
- [19] José Eduardo Marques Pessanha, Silvana Tecles Brandão, Maria Cristina Mattos Almeida, Maria da Consolação Magalhães Cunha, Ivan Vieira Sonoda, Adelaide Maria Bessa, and José Carlos Nascimento. Ovitrap surveillance as dengue epidemic predictor. *Journal of Health & Biological Sciences*, 2(2):51–56, 2014.
- [20] José Eduardo Marques Pessanha. Onde está wally? ou onde se esconde o aedes aegypti. *Boletim Epidemiológico*, X(4):26, 2007.

- [21] Duc Nghia Pham, Tarique Aziz, Ali Kohan, Syahrul Nellis, Jing Jing Khoo, Dickson Lukose, Sazaly AbuBakar, Abdul Sattar, Hong Hoe Ong, et al. How to efficiently predict dengue incidence in kuala lumpur. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, pages 1–6. IEEE, 2018.
- [22] Kazi Mizanur Rahman, Yushuf Sharker, Reza Ali Rumi, Mahboob-Ul Islam Khan, Mohammad Sohel Shomik, Muhammad Waliur Rahman, Sk Masum Billah, Mahmudur Rahman, Peter Kim Streatfield, David Harley, et al. An association between rainy days with clinical dengue fever in dhaka, bangladesh: findings from a hospital based study. *International Journal of Environmental Research and Public Health*, 17(24):9506, 2020.
- [23] Sandali Raizada, Shuchi Mala, and Achyut Shankar. Vector borne disease outbreak prediction by machine learning. In *2020 International conference on smart technologies in computing, electrical and electronics (ICSTCEE)*, pages 213–218. IEEE, 2020.
- [24] Sujit Sahu. *Bayesian modeling of spatio-temporal data with R*. Chapman and Hall/CRC, 2022.
- [25] Andre Ricardo SALATA and Marcelo Gomes RIBEIRO. Boletim desigualdade nas metrópoles. <https://www.observatoriodasmetropoles.net.br/> note = Disponível em: Observatório das Metrópoles e PUCRS. Acesso em: 10 jun. 2024, 2024.
- [26] Ignacio Sanchez-Gendriz, Matheus Diniz, AD Doria Neto, Rodrigo Moreira Pedreira, Ion de Andrade, and RA de Medeiros Valentim. Deep learning-based ovitrap spatial dynamics analysis for arbovirus vector monitoring. *XVI Brazilian Conference on Computational Intelligence*, 2023.
- [27] Dhiman Sarma, Sohrab Hossain, Tanni Mittra, Md Abdul Motaleb Bhuiya, Ishita Saha, and Ravina Chakma. Dengue prediction using machine learning algorithms. In *2020 IEEE 8th R10 humanitarian technology conference (R10-HTC)*, pages 1–6. IEEE, 2020.
- [28] Juan M Scavuzzo, Francisco Trucco, Manuel Espinosa, Carolina B Tauro, Marcelo Abril, Carlos M Scavuzzo, and Alejandro C Frery. Modeling dengue vector population using remotely sensed data and machine learning. *Acta tropica*, 185:167–175, 2018.
- [29] Olivia Lang Schultes, Maria Helena Franco Moraes, Maria da Consolação Magalhães Cunha, Andréa Sobral, and Waleska Teixeira Caiaffa. Spatial analysis of dengue incidence and aedes aegypti ovitrap surveillance in belo horizonte, brazil. *Tropical Medicine & International Health*, 26(2):237–255, 2021.
- [30] Roberto CSNP Souza, Renato M Assunção, Derick M Oliveira, Daniel B Neill, and Wagner Meira Jr. Where did i get dengue? detecting spatial clusters of infection risk with social network data. *Spatial and spatio-temporal epidemiology*, 29:163–175, 2019.
- [31] Lucas M Stolerman, Pedro D Maia, and J Nathan Kutz. Forecasting dengue fever in brazil: An assessment of climate conditions. *PloS one*, 14(8):e0220106, 2019.
- [32] Sediyama GC. Vianello RL, Pessanha JEM. Previsão de ocorrência dos mosquitos da dengue em belo horizonte com base em dados meteorológicos. 2006.
- [33] Naizhuo Zhao, Katia Charland, Mabel Carabali, Elaine O Nsoesie, Mathieu Maheu-Giroux, Erin Rees, Mengru Yuan, Cesar Garcia Balaguera, Gloria Jaramillo Ramirez, and Kate Zinszer. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in colombia. *PLoS neglected tropical diseases*, 14(9):e0008056, 2020.