



UNIVERSIDADE FEDERAL DE MINAS GERAIS

Trabalho de Conclusão de Curso

Aluno
Pedro Barbosa Bahia

Orientadores
Frederico Coelho e Renato Assunção

Janeiro de 2025

Conteúdo

1	Introdução	2
2	Pesquisa Bibliográfica	2
2.1	Área de Estudo	2
2.2	Ovitrampas	2
2.3	Breve Histórico	3
2.4	Estado da Arte	5
3	Descrição dos Dados	5
3.1	Ovitrampas	5
3.2	Variáveis exógenas	17
4	Metodologia	18
4.1	Seleção de entradas	22
4.2	Modelos de Aprendizado de Máquina	23
4.3	Métricas	23
4.4	Seleção de Features	23
4.5	Resíduos	25
4.5.1	Classificação de Presença	25
4.5.2	Régressão	25
4.5.3	Cross Validation	27
5	Conclusão	27
6	Apêndice	27
6.1	Tratamento de Valores Inconsistente	27
6.2	Modelos iniciais	27
7	Bibliografia	28

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. (18) 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.(28)

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 (7), 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 (24).

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 (6). Ela apresenta temperatura média anual de 22,1 °C e precipitação acumulada anual média de 1.578,3 mm (9) 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 (9).

2.2 Ovitrampas

Ovitrampas, Figura (18), 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 (18)

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 (19).

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. (19)

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 (31) quanto embasam novos modelos. Por exemplo, (17) 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

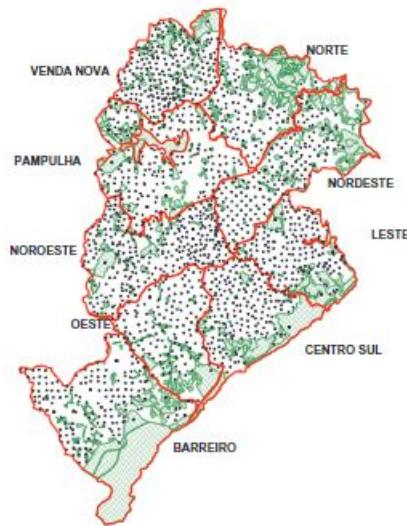


Figura 2: Grade de ovitrampas em Belo Horizonte (18)

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. (18)

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. (14) 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. (14)

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.(30; 1; 32) Porém, parâmetros relacionados a variáveis obtidas por sensoriamento remoto(12), dados de redes sociais (29), consultas em ferramentas de busca (16) e, principalmente, dados relacionados ao monitoramento do vetor (2; 10; 15) também são ocasionalmente utilizados quando disponíveis.

Em relação às técnicas, os modelos de Poisson (5) e modelos de média móvel (ARIMA, SARIMA, ARIMAX) (3; 21; 20) são amplamente utilizados para previsão de séries históricas. Entretanto, constata-se o crescente uso de Redes Neurais Artificiais(22; 11), principalmente Long Short Term Memory (LSTM) (4), Máquinas de Vetores de Suporte (SVM) (27) e modelos baseados em Árvores de Decisão (26).

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 (25) 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. (28)

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 3.



Figura 3: Mapa com a localização das armadilhas analisadas

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 incorreta de datas de coleta. Mais detalhes sobre o tratamento dos dados problemáticos encontram-se no Apêndice !!!!!!!!.

Com os dados corrigidos, calculou-se a diferença entre as datas de instalação e de coleta de cada placa, ou seja, seu tempo de exposição (Figura 4). Como esperado, as amostras se aglutinaram consistentemente entre 6 e 8 dias, consistindo em 99% das amostras disponíveis. Outro valor avaliado foi a diferença entre as datas de coletas de placas da mesma armadilha, equivalente à taxa amostral da armadilha (Figura 5). Apesar das intempéries que podem acarretar ausência de amostras na base de dados, como seu descarte no processamento inicial, 80% das amostras distam-se entre 13 e 15 dias, com moda clara em 14 dias. Um segundo grupo relevante, entretanto, equivalente a 8% do total, dista por volta de 28 dias. Isso seria explicado pelo fato de, consistentemente ao longo dos anos, uma amostra de cada armadilha não ser coleta no mês de dezembro e outra no mês do carnaval, conforme

explicado !!!!!!!!. Essa característica pode ser verificada também na contagem de armadilhas por mês na Figura 6, em que o número de amostras nos meses epidemiológicos 7 (dezembro) e 9 (fevereiro) é aproximadamente a metade dos demais.

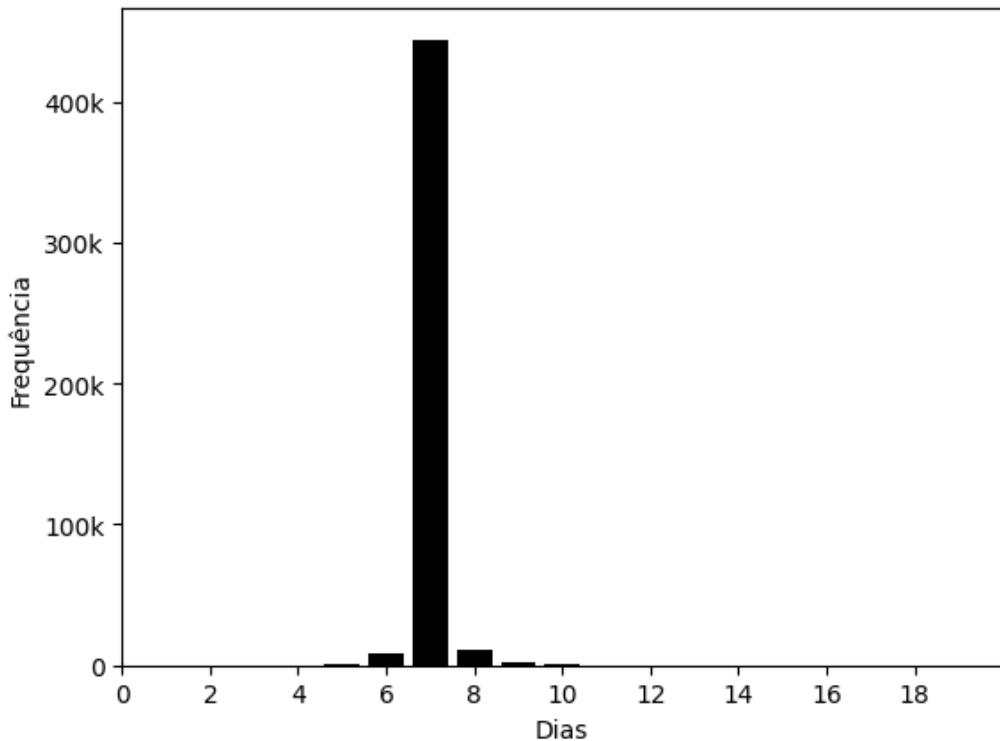


Figura 4: Histograma do tempo de exposição das placas

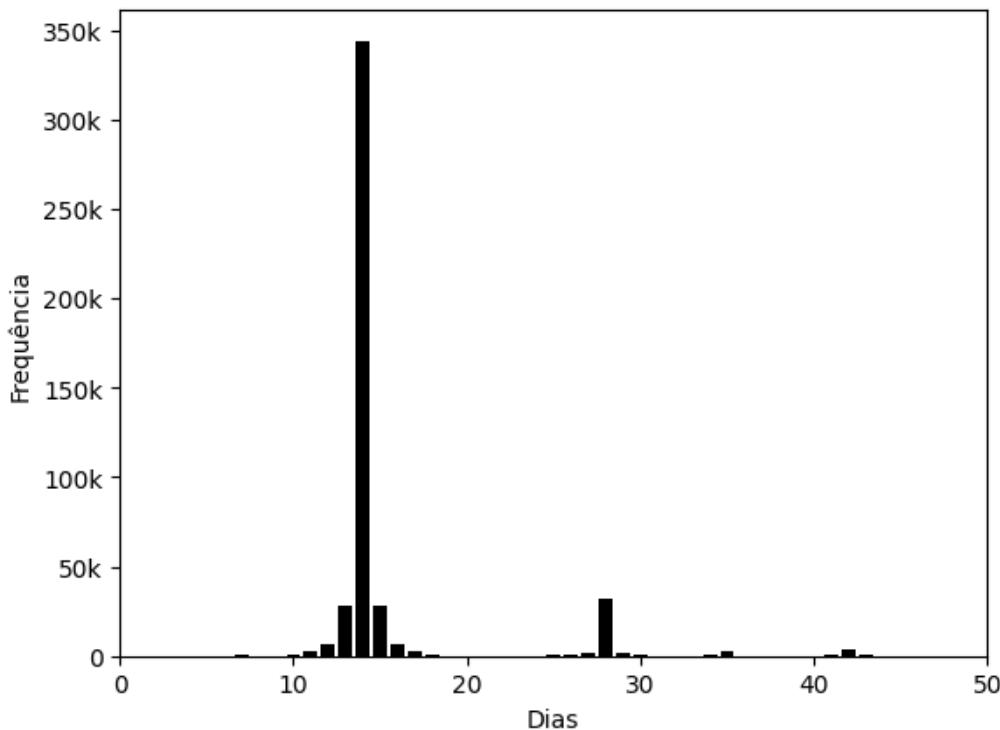


Figura 5: Histograma da frequência de amostragem das armadilhas

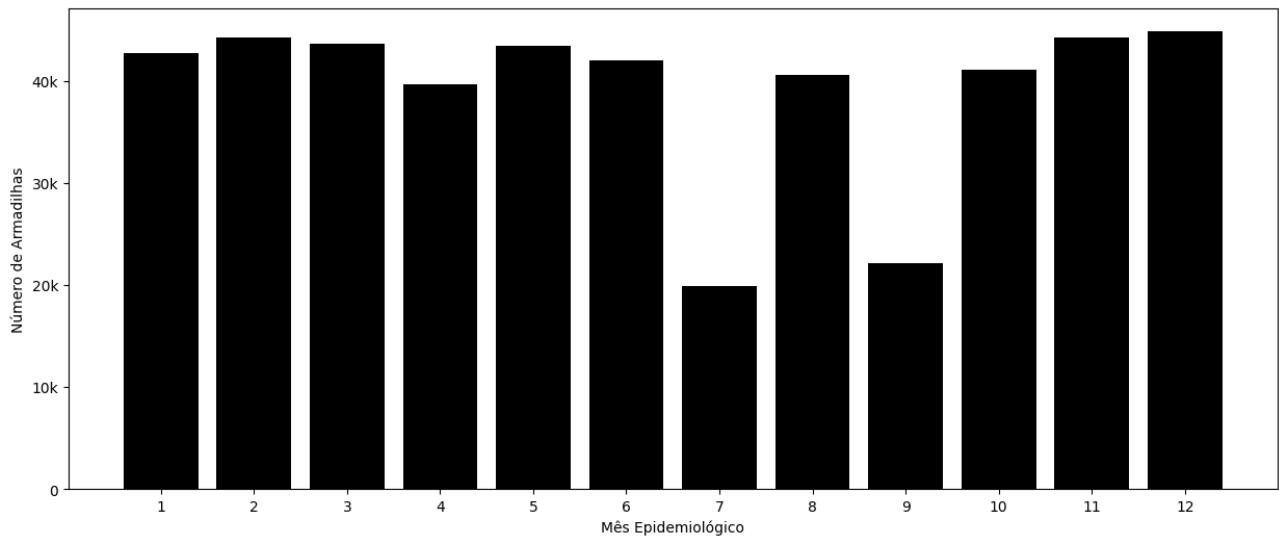


Figura 6: Número de placas por mês epidemiológico

Além da variação mensal, o número de placas apresenta variação anual, referente ao processo gradual de implementação da política pública e na reestruturação da malha. Conforme Figura 7, o número de coletas demonstra tendência de crescimento anual, à parte aos anos epidemiológicos de 2011_12 e 2024_25 que estão incompletos na base utilizada. Além disso, no mapa da Figura 8, em que o raio do círculo posicionado no local de cada armadilha é proporcional à contagem de placas associada a ela. O número médio de amostras por armadilha é de aproximadamente 260, enquanto o número máximo é 311. Entretanto, nota-se número considerável de armadilha com raio é menor que a média das demais. Isso indica, em conformidade com o exposto em !!!!!!!!!!!!!!!, a implementação de novas armadilhas nesses locais em anos recentes.

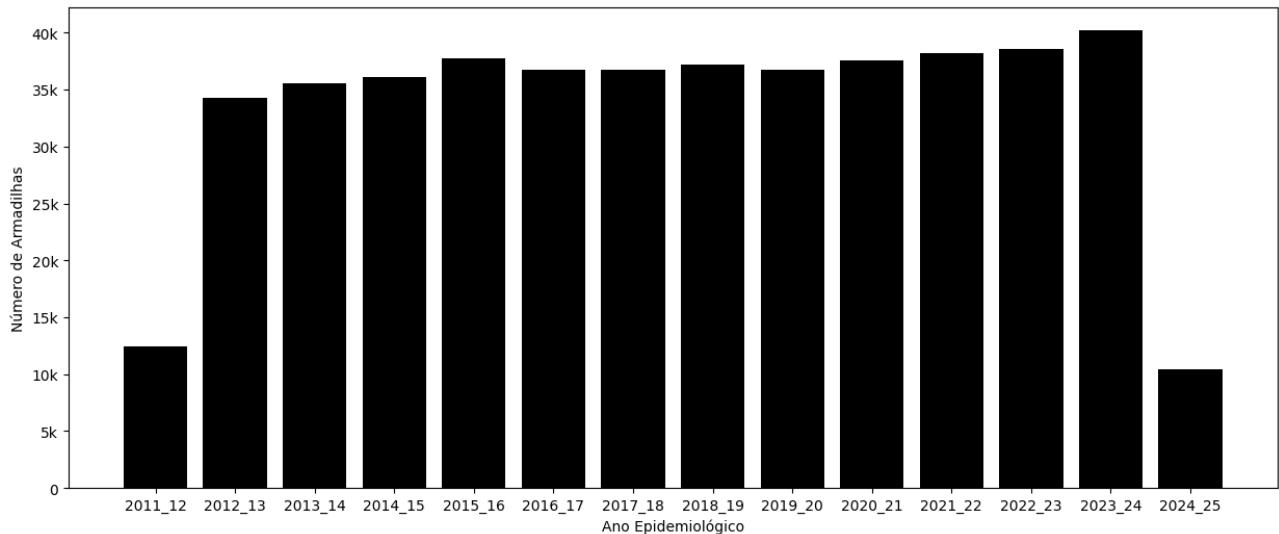


Figura 7: Número de placas por ano epidemiológico

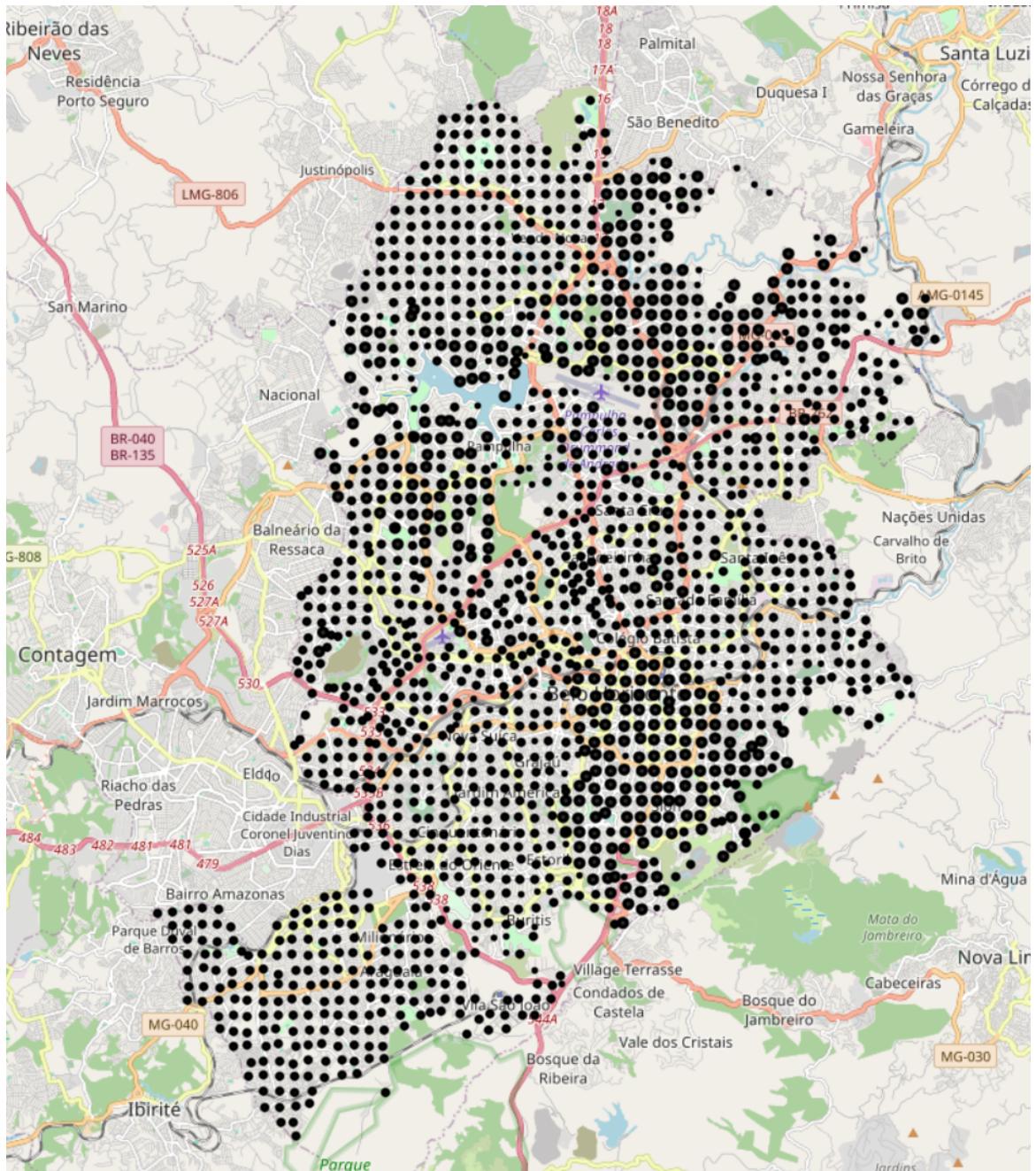


Figura 8: Número de placas por localidade

Dada a diferença na quantidade de amostras disponíveis por ano e por armadilha, a soma do número de ovos não seria um critério relevante para comparação e descrição global. Por isso, a média de ovos por número de placas foi utilizada na verificação de tendências nos dados. A variação mensal na média de ovos (Figura 9), por exemplo, explicita a componente sazonal associada às estações do ano. Os meses de verão (7, 8 e 9), cujo regime de chuvas contribui para a reprodução e proliferação do *Aedes aegypti*, apresentam valores médios consideravelmente maiores que os meses de inverno (2, 3 e 4). Desse modo, os efeitos do baixo número de amostras em dezembro e fevereiro (7 e 9) é reduzido. De modo análogo, a média de ovos por ano epidemiológico da Figura 10 é agnóstica ao número de amostras disponíveis e condiz com a ocorrência de anos de epidemia (2012_13, 2018_19, 2022_23, 2023_24,!!!!!!!!!!!!!!!). Pontua-se apenas a discrepância entre os valores médios dos anos de 2011_12 e 2024_25. Apesar do número de amostras equivalente em ambos os anos, as amostras disponíveis do primeiro são referentes aos últimos meses epidemiológicos do ano, enquanto as amostras do último são referentes aos primeiros meses.

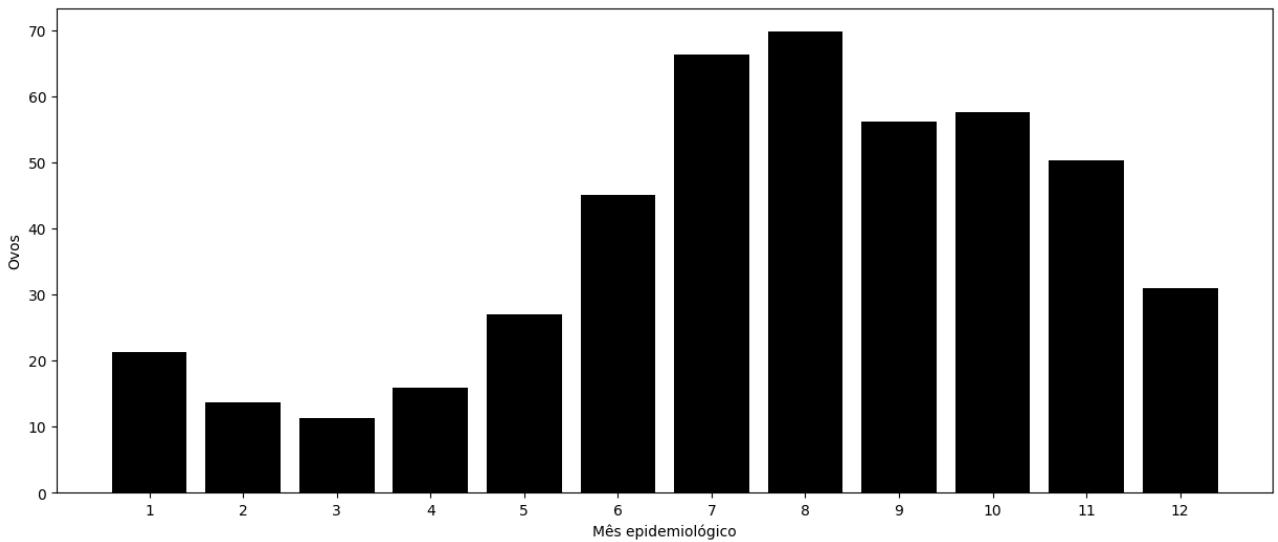


Figura 9: Média de ovos por mês epidemiológico

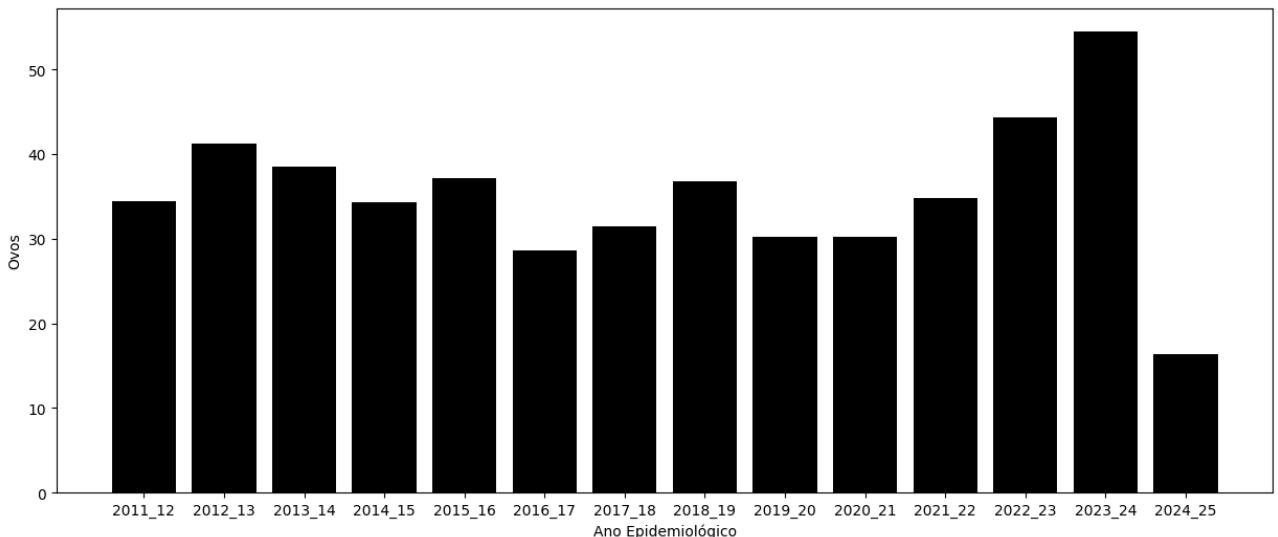


Figura 10: Média de ovos por ano epidemiológico

De forma análoga, a Tabela 2 apresenta o número de placas de cada categoria, a soma dos ovos pertencentes a cada uma delas e suas respectivas médias. Apesar de apenas 5.1% das placas pertencer à Categoria A1, ela compreende 13.5% de todos os ovos coletados, com uma média duas vezes maior que a categoria A2. De modo semelhante, armadilhas da categoria B compreendem apenas 3.0% do total de ovos, enquanto representam 12.8% das amostras.

Categoria	Número de Placas	Soma dos Ovos	Média
A1	23.7K	2.3M	97.0
A2	291.4K	12.4M	42.4
M	93.5K	1.9M	20.0
B	60.0K	0.5M	9.0

Tabela 2: Placas e Ovos por Categorias

Partindo desta distoância entre as categorias de armadilhas e iniciando uma análise com maior resolução do espacial, o mapa 11 foi gerado para identificar a dinâmica espacial das armadilhas. A cada

armadilha foi atribuída uma cor de acordo com a categoria à qual ela pertencia e um raio proporcional à média de ovos coletadas nela. A partir dele identifica-se uma clara estrutura espacial, com armadilhas da categoria B concentrando-se no sudeste da cidade, enquanto armadilhas da categoria A1 encontram-se em maior quantidade em porções centrais e setentrionais. Ademais, alguns agrupamentos de armadilhas dessas duas categorias podem ser identificados.

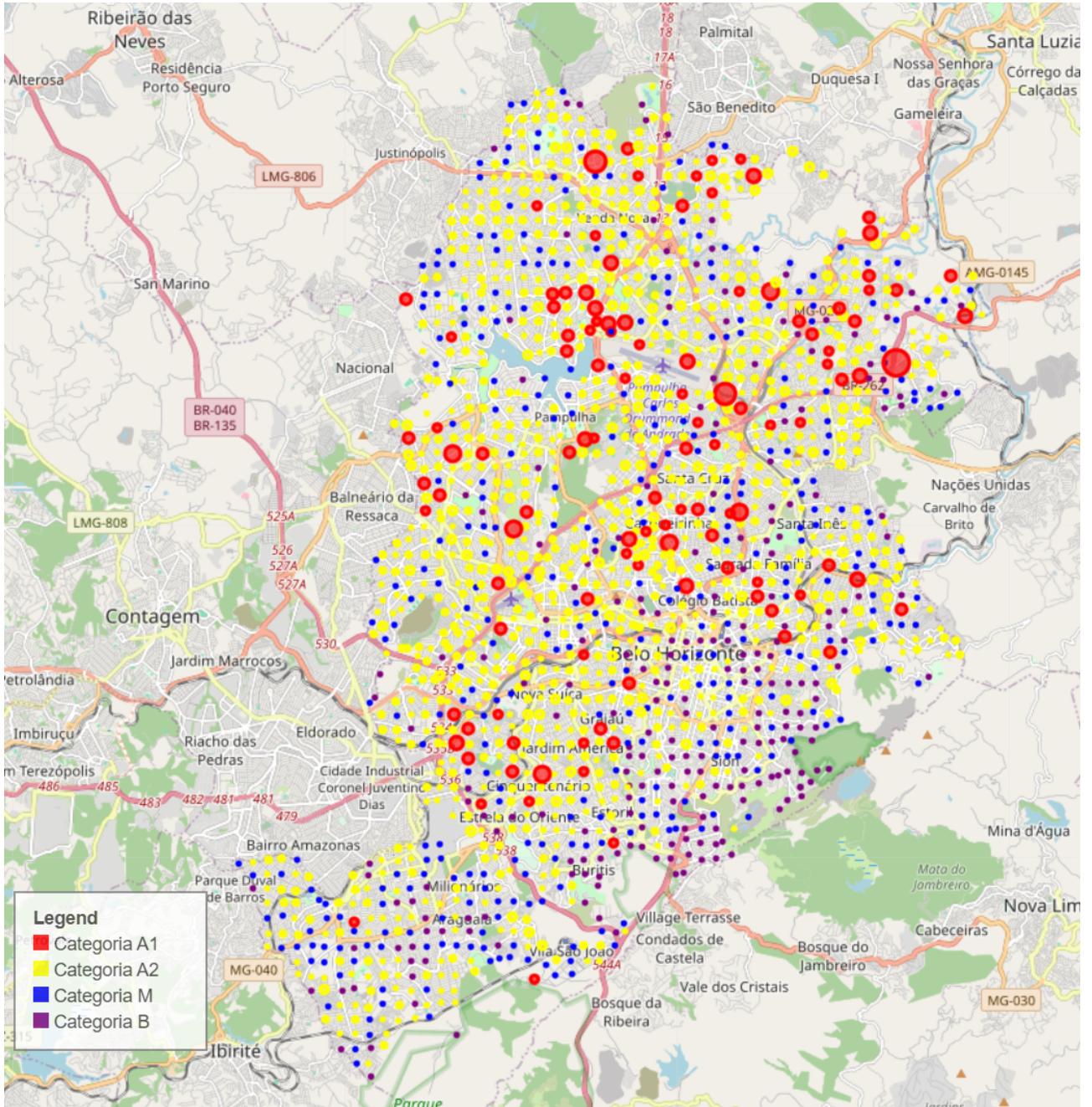


Figura 11: Mapa das armadilhas por classe

No sentido de continuar as análises de maior resolução, as amostras individuais foram analisadas. Com uma média de 36 ovos por coleta, a contagem de ovos estende-se de 0 a 4227. A mediana, também 0, indica quantidade considerável de placas sem a presença de ovos. O histograma 12 ilustra a distribuição das ovitrampas. O histograma (12a) exprime a prevalência de placas vazias, além da amplitude do número de ovos encontrados. Para melhorar a visibilidade, as amostras sem ovos foram omitidas na imagem (12b) e na imagem (12c) foram contabilizados apenas armadilhas entre 2 e 1000 ovos. A média para amostras não nulas aumentou para 75 ovos, enquanto a mediana para 48.

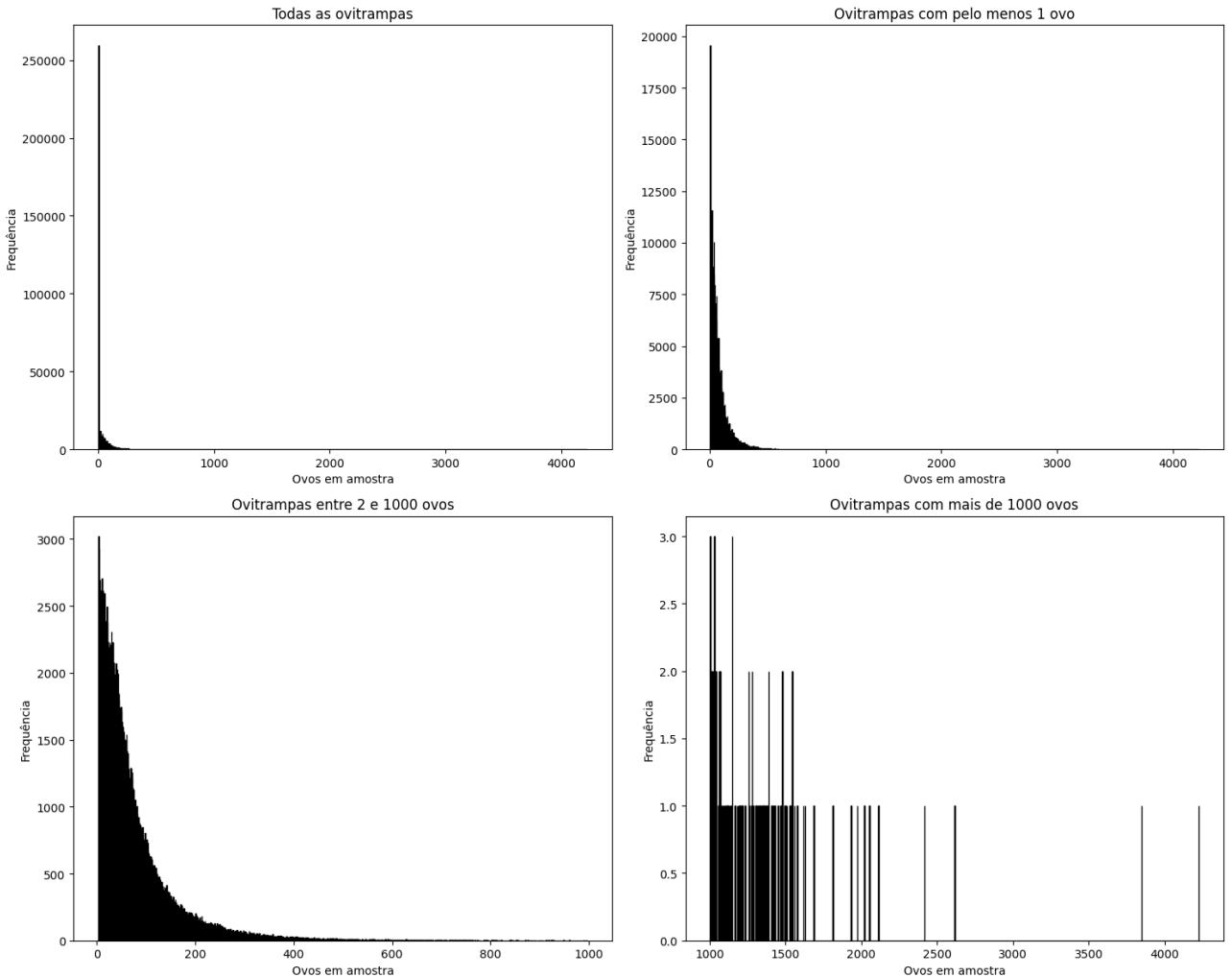


Figura 12: Histogramas das quantidades de ovos por ovitrapa. Da esquerda para a direita, de cima para baixo: (a) todas as ovitrampas; (b) ovitrampas com pelo menos um ovo; (c) ovitrampas entre 2 e 1000 ovos; (d) ovitrampas com mais de 1000 ovos

Além da abundância de amostras com apenas um ovo, o histograma apresenta decaimento aparentemente exponencial, porém seguido de uma cauda pesada, incomum em distribuições exponenciais, detalhada melhor na imagem (12d) em que apenas amostras com mais de 1000 ovos são apresentadas. A partir deste valor, são frequentes contagens únicas, com estas aparições tornando-se esparsas à medida que os valores aumentam. Tais caudas pesadas são características de um grupo de distribuições do qual um dos representante mais conhecidos é a distribuição Pareto, que pode ser identificada a partir de um gráfico log-log, ou seja, do logaritmo da frequência pelo logaritmo do valor de interesse. A Figura 13, gerada a partir de !!!!!!!!!!!!!!!!, ilustra este comportamento: um alinhamento linear seguido de um espalhamento horizontal dos valores. Já na Figura 14, os gráficos log-log de todas as ovitrampas (14a) e das ovitrampas com valores maiores que 100 (14b) são apresentados. Esta clivagem foi escolhida pois, a partir deste valor, o gráfico 14a começa a se assemelhar ao 13, com uma relação aproximadamente linear seguida do espalhamento horizontal. Este valor, portanto, será considerado o início aproximado da cauda pesada da distribuição de ovitrampas.

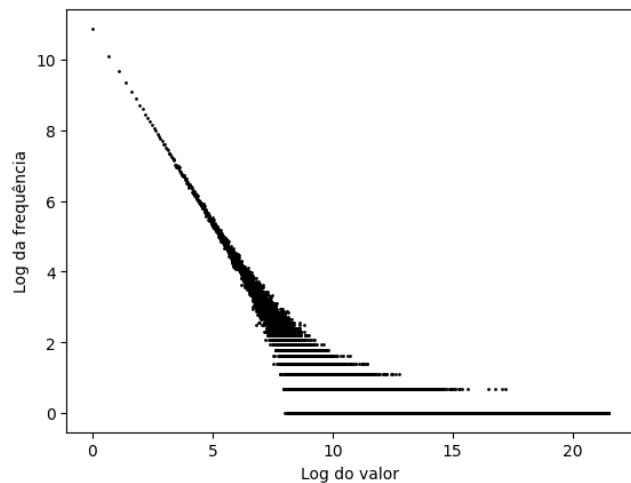


Figura 13: Gráfico log-log de uma distribuição Pareto artificial

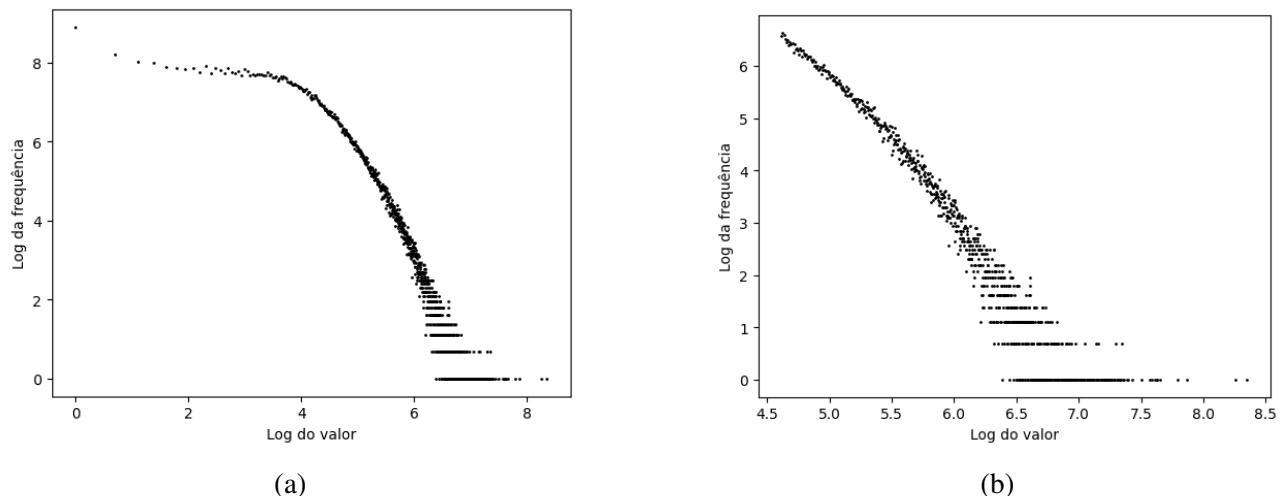


Figura 14: Gráfico log-log da distribuição de ovitrampas. Da esquerda para a direita: (a) todos as contagens de ovos são incluídas; (b) incluídas contagens a partir de 100

Outra característica de interesse da distribuição avaliada é a quantidade de amostras nulas. Como demonstrado, pelo menos 50% das amostras totais estão vazias. O gráfico 15 apresenta a porcentagem de ovitrampas vazias por categoria ao lado da porcentagem das ovitrampas com valores não nulos, para efeito de comparação com o conjunto de amostras completo. É perceptível que a quantidade de placas vazias reduz à medida que as armadilhas são classificadas como mais propensas a terem ovos. Desse modo, o baixo número médio de ovos coletados em uma localidade provavelmente se deve ao aumento do número de coletas com ausência de ovos. O gráfico 16, por sua vez, ilustra a porcentagem de zeros por ano e mês epidemiológicos. Em concordância com os resultados anteriores, constata-se que o número de amostras nulas aumenta nos meses e anos em que a média de ovos é mais alta.

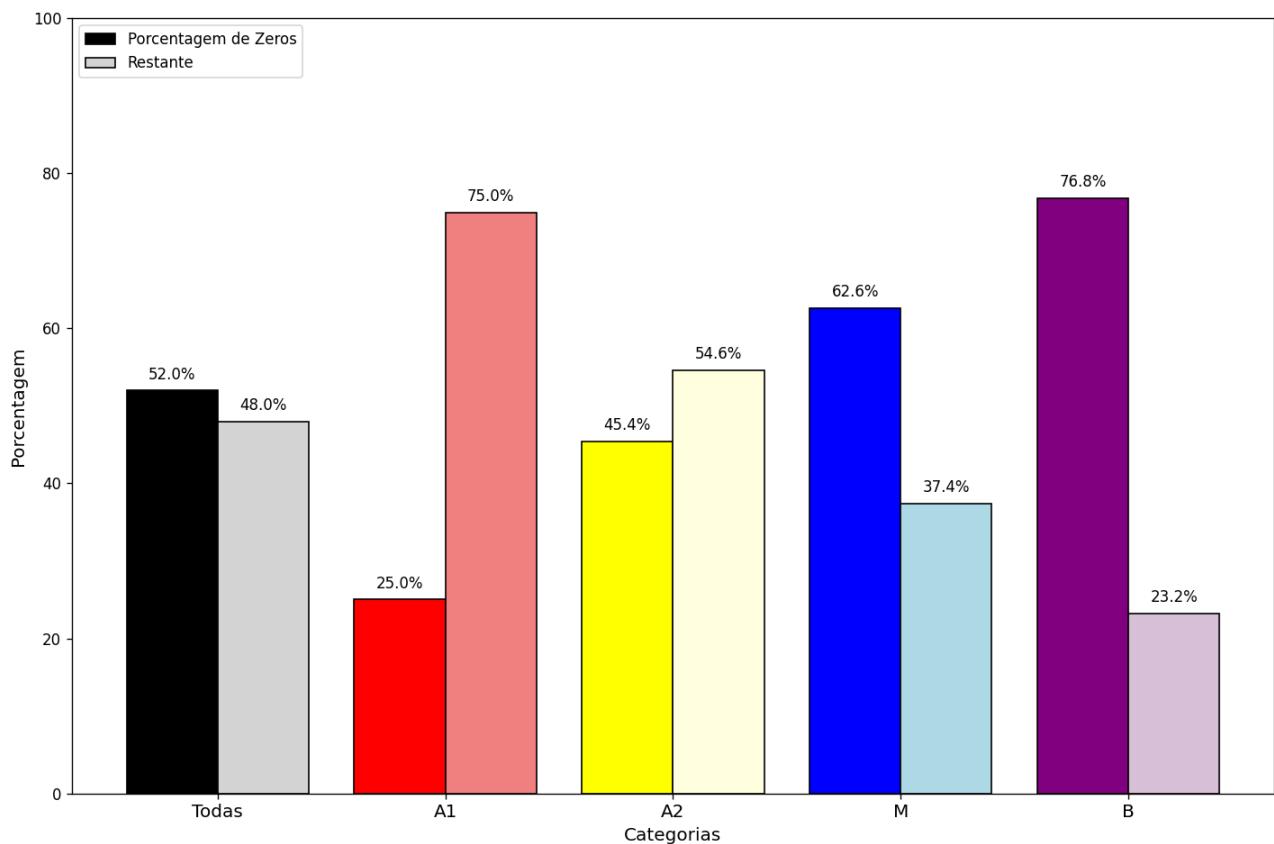
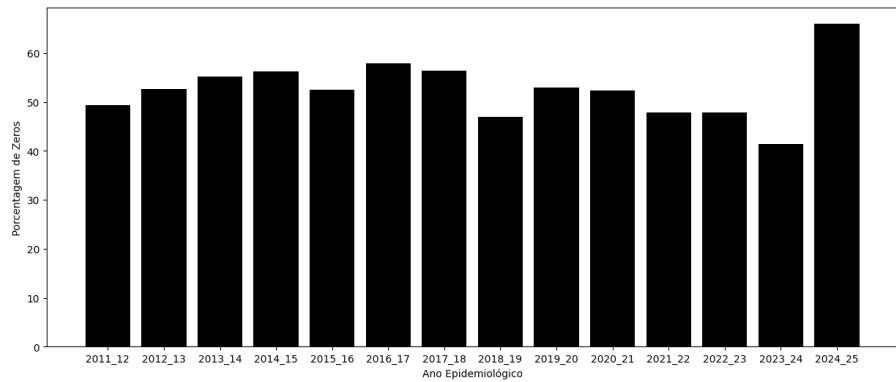
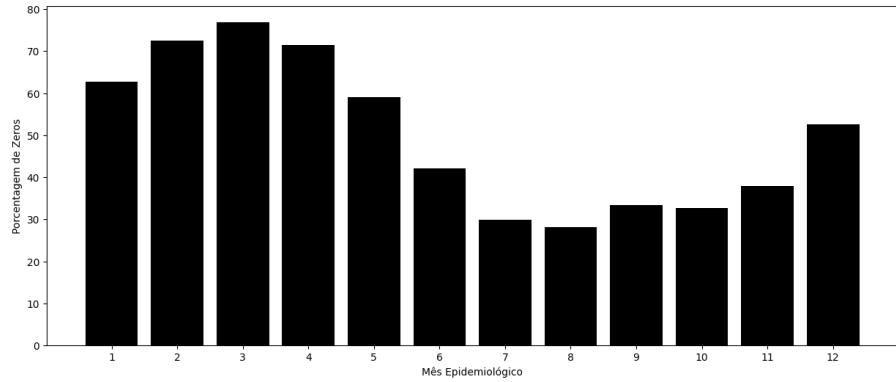


Figura 15: Porcentagem de ovitrampas vazias em todo o banco de dados e por categoria



(a)



(b)

Figura 16: Gráficos com a porcentagem de amostra vazias por (a) ano epidemiológico e (b) mês epidemiológico

Por fim, para uma análise mais detalhada da estrutura temporal, as médias das amostras coletadas foi discriminada por ano e mês epidemiológicos, como apresentado na Figura 17. De fato, constata-se componente sazonal relevante associada ao mês do ano, já que as tendências verificadas em um mês epidemiológico se repetem consistentemente ao longo dos anos. Essa análise aplicada às quatro classes de armadilhas corroboram com a mesma conclusão, conforme ilustrado em 18.

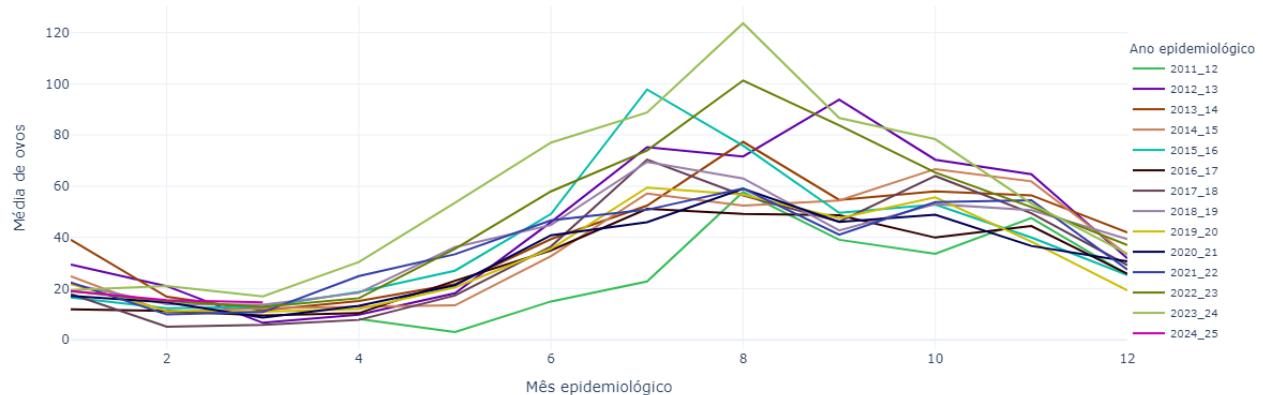


Figura 17: Série histórica da média de ovos separados por ano

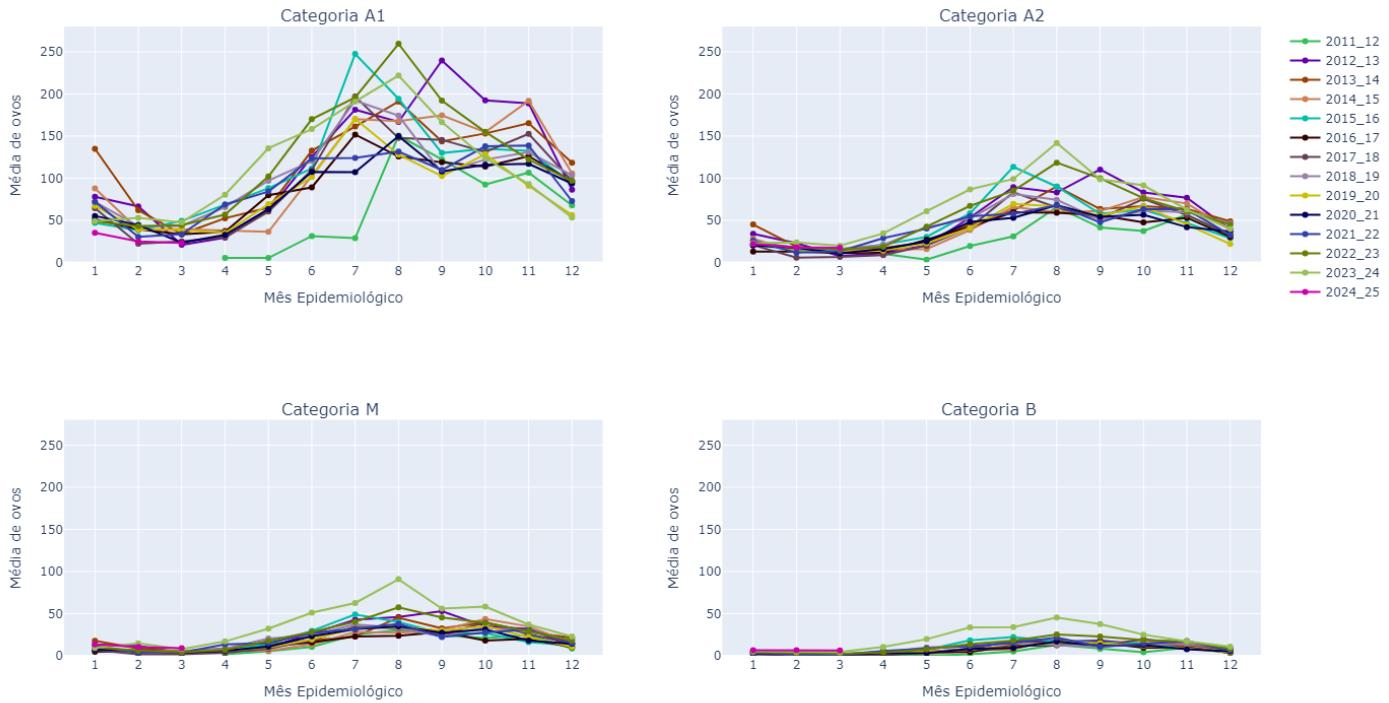


Figura 18: Série histórica da média de ovos separados por ano em cada classe de armadilhas

3.2 Variáveis exógenas

Além dos dados das ovitrampas fornecidos pela PBH, no presente estudo serão utilizadas as variáveis exógenas temperatura, umidade e pluviometria, devida a alta correlação com o número de ovos !!!!!!!!!!!!!!! ref. Tais dados têm como origem cinco estações meteorológicas do Instituto Nacional de Meteorologia (INMET) (8) e duas estações do Departamento de Controle do Espaço Aéreo (DECEA) [!!!!!!!] e foram obtidos por meio ao acesso à base de dados das próprias instituições [ref 1-4!!!!!!!]

As estações cobrem a região metropolitana de Belo Horizonte em uma malha não regular, conforme mapa da Figura 19, coletando dados por períodos com início e finais distintos, em alguns casos não cobrindo o período de coleta das ovitrampas. Além disso, os dados dos pontos de coleta foram registrados com taxas amostrais variáveis nos três regimes: a cada hora, três vezes por dia (às 00h, 12h e 18h) e quatro vezes por dia (às 00h, 06h, 12h e 18h)

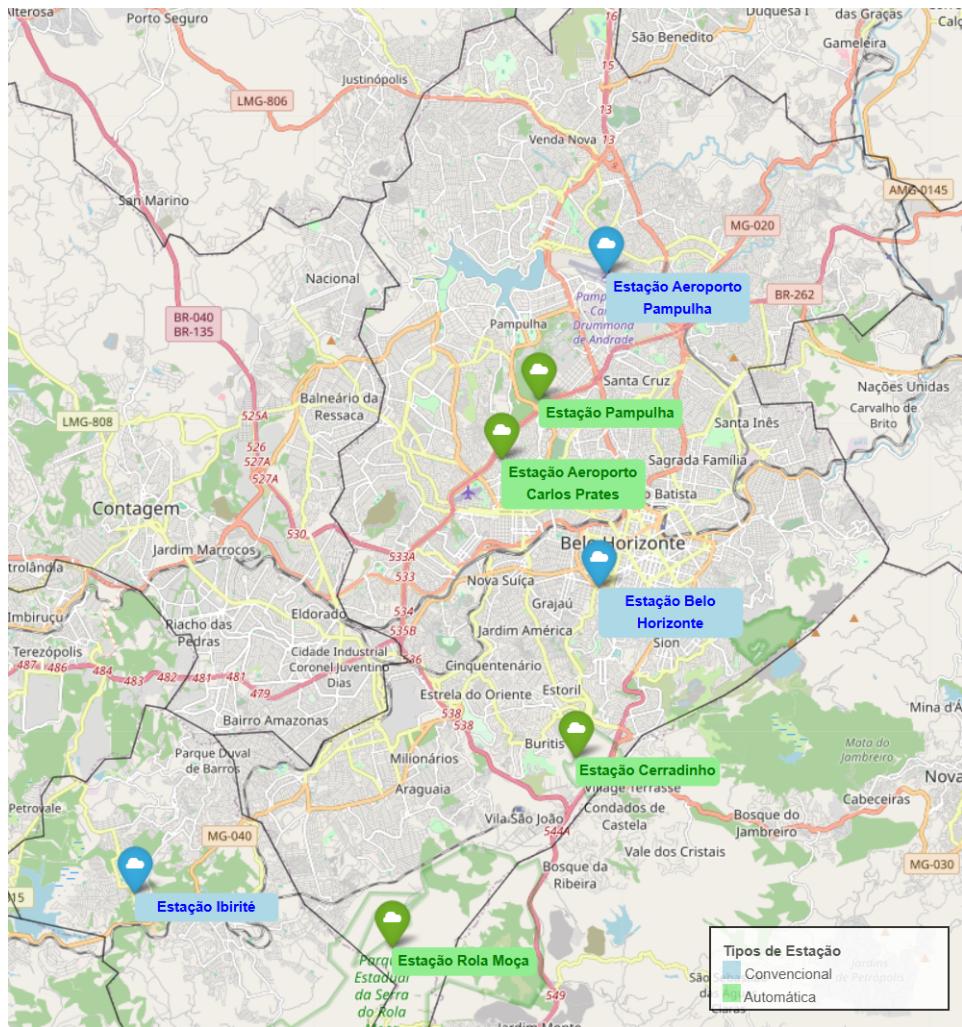


Figura 19: Localização das estações meteorológicas e pluviômetros na região metropolitana de Belo Horizonte (8)

4 Metodologia

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 entomológicos. 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.

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.

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 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

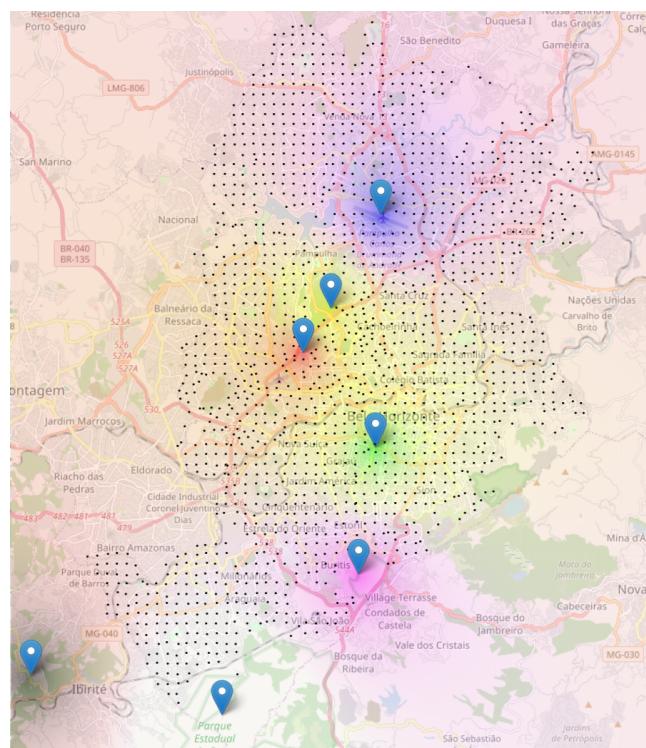


Figura 20: Caption

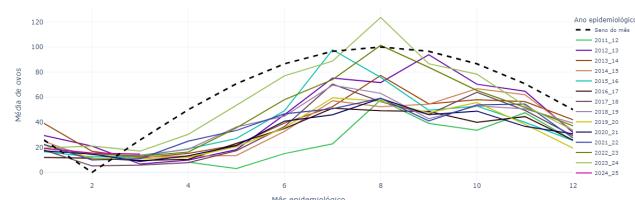


Figura 21: Caption

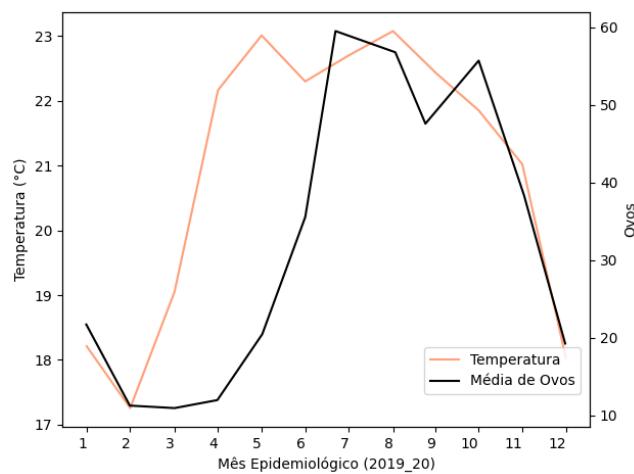


Figura 22: Caption

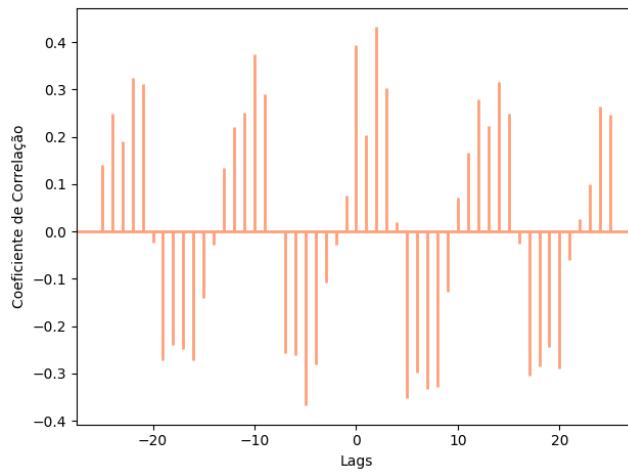


Figura 23: Caption

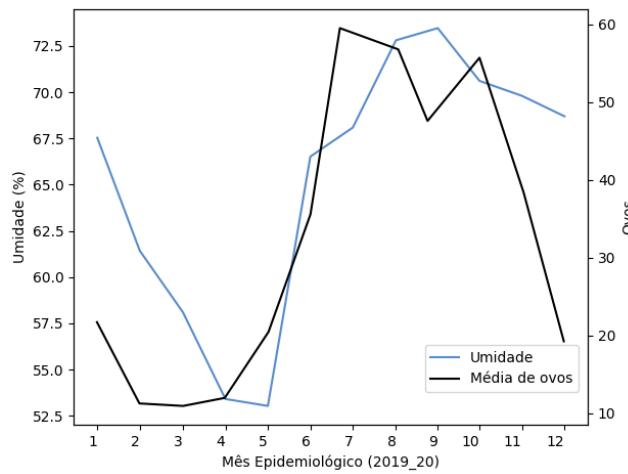


Figura 24: Caption

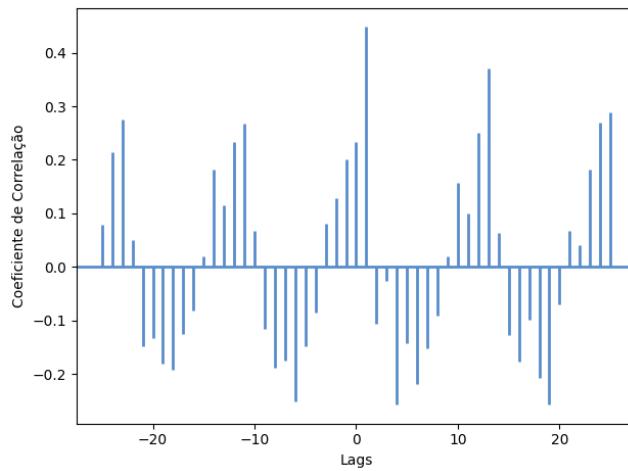


Figura 25: Caption

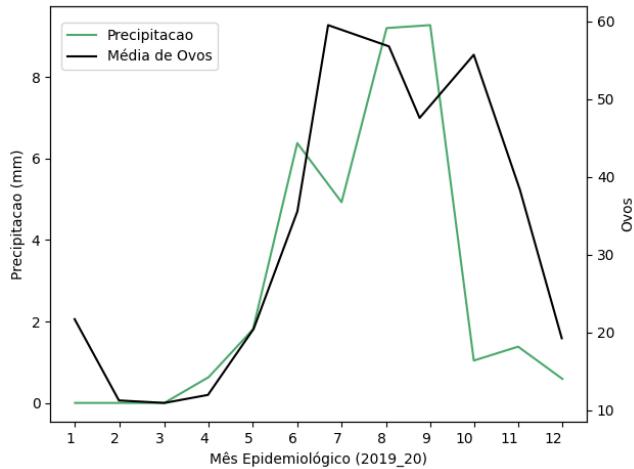


Figura 26: Caption

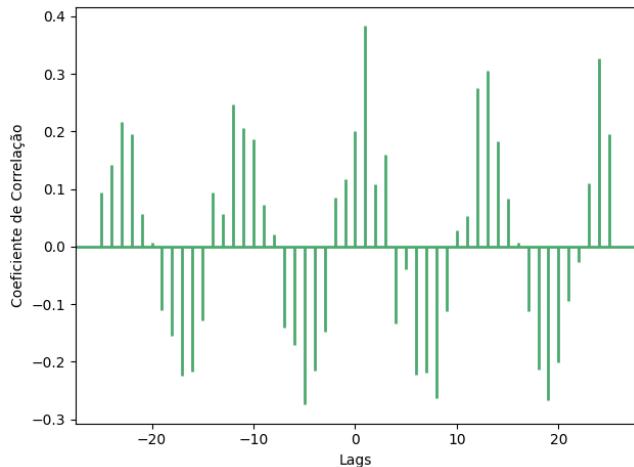


Figura 27: Caption

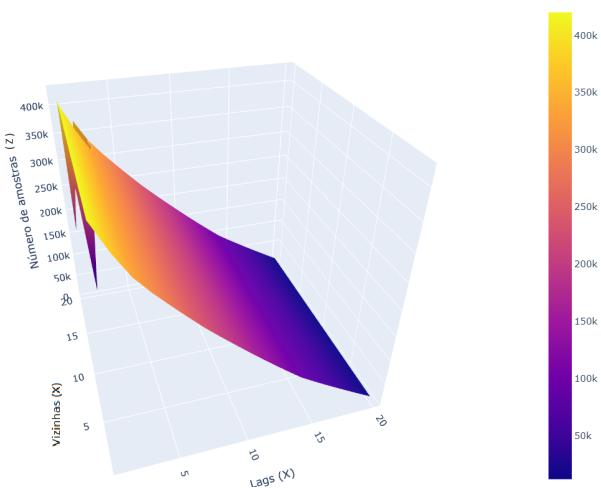


Figura 28: Caption

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	Sete da semana epidemiológica de instalação da amostra
semeipi2	Quadrado da sete da semana epidemiológica de instalação da amostra
sin_semeipi	Seno da sete 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

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 (23), será utilizado como medida da correlação da ovitrampa analisada com os K vizinhos mais 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 EC as variáveis exógenas categóricas.

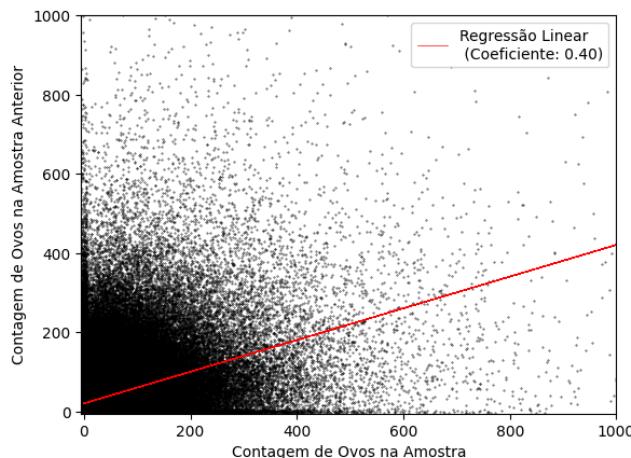


Figura 29: Caption

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 compreende 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 (13). 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.

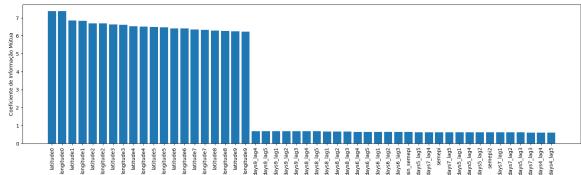
4.3 Métricas

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. (13) (25)

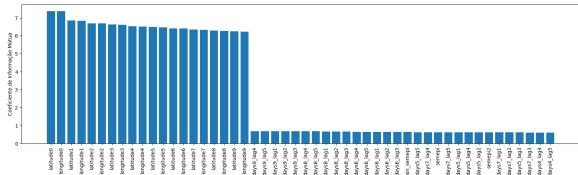
Dois modelos serão utilizados como referência para avaliar o desempenho dos demais. Um modelo ingênuo, que repete o valor anterior da armadilha como predição, e um modelo SARIMAX, comumente utilizado para esse fim (13). SARIMAX é um modelo estatístico utilizado em previsões de séries temporais estacionárias, 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
Curiosamente, lat e long são muito correlacionados



(a) First figure description.



(b) Second figure description.

Figura 30: Overall caption for the figures.

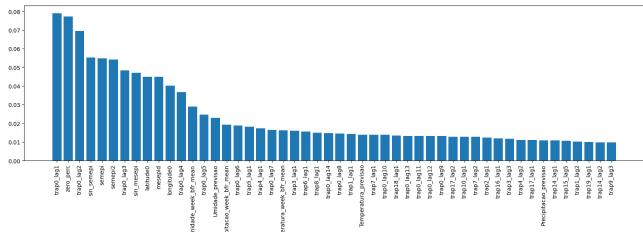


Figura 31: Caption

[‘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, eg' 38.7(35.4) 'svr' 42.3(36.9) –

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) diferenamaxima0.1'svm'59.3(64.6)diferenamaxima14 - melhor : 70.0'catboost'70.8(73.3)diferenamaxima0.33_classes : perc-ero'logistic3c'72.5(75.6)'Naive3c'65.2(68.1)'mlp3c'72.6(76.0)30epocassemmudanca, de 10

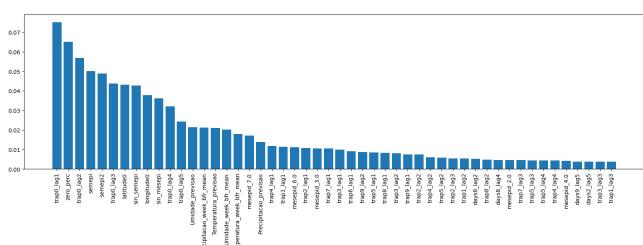


Figura 32: Caption

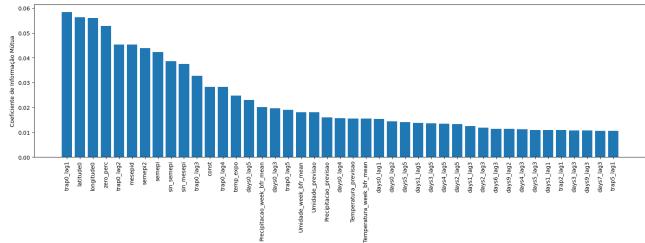


Figura 33: Caption

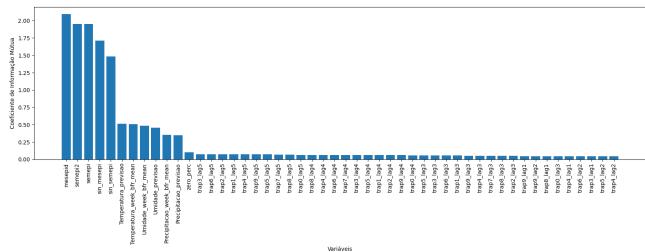


Figura 34: Caption

5) *naloss, adam, relu, random_forest3c'72.5(78.2)'svm3c'71.6(74.7)'catboost3c'72.7(76.0)'linear3c'59.3(64.7)*

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

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

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, 30.2772612023, 40.3033852024, 50.294327

Cat A1 0.267086 A2 0.309356 B1 0.224403 M1 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, 320.543952023, 424.8956272024, 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 sínais 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.

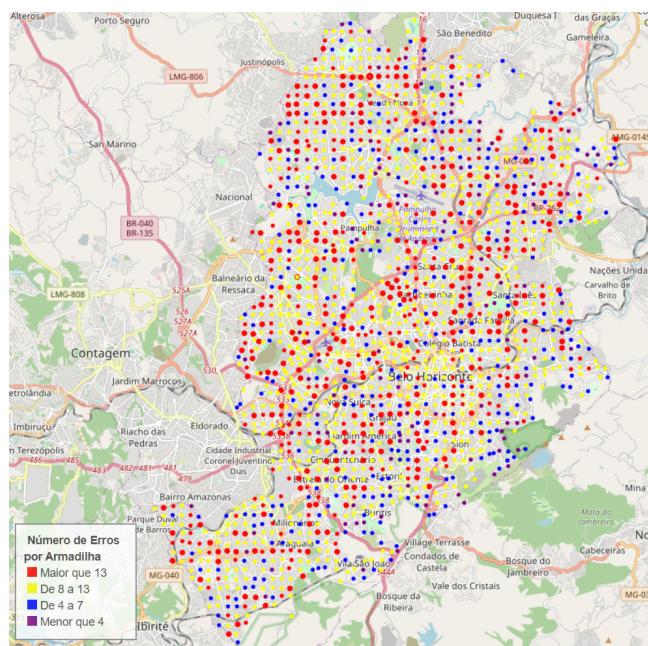


Figura 35: Caption

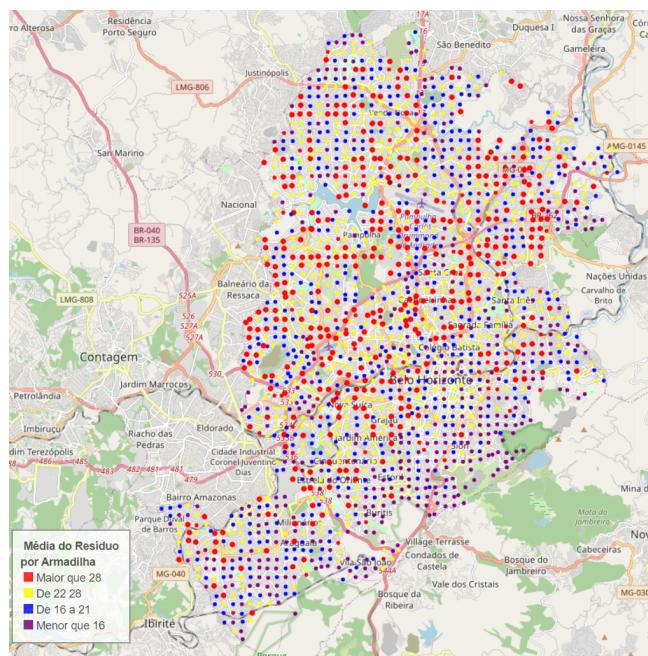
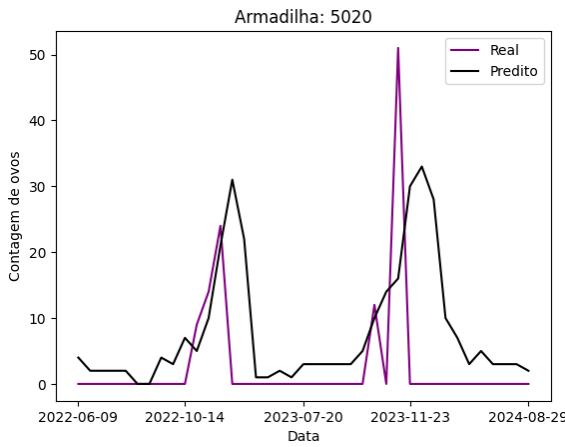
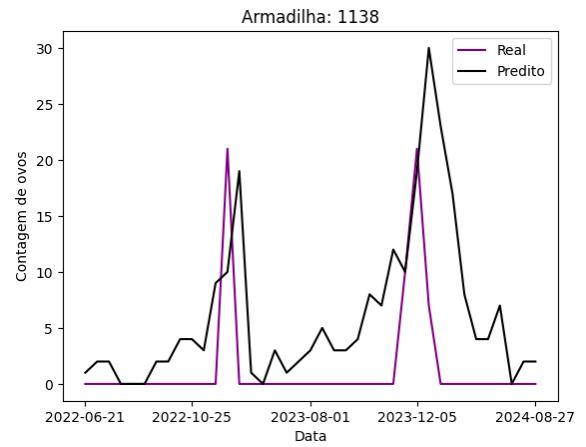


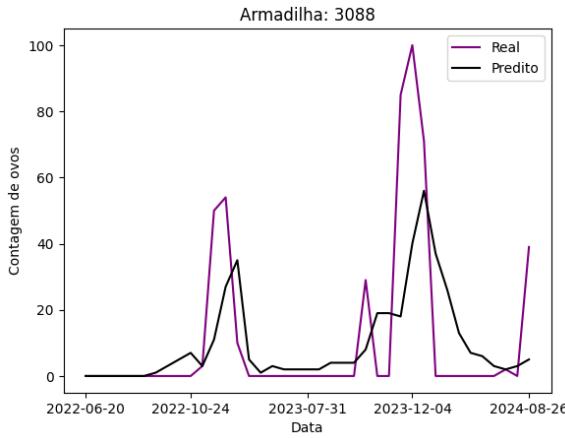
Figura 36: Caption



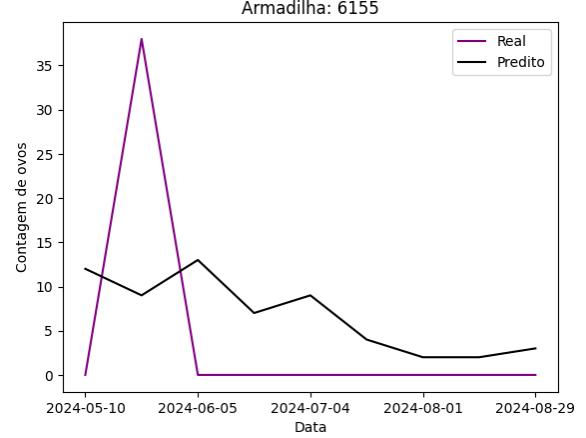
(a) First figure description.



(b) Second figure description.



(c) First figure description.



(d) Second figure description.

Figura 37: Overall caption for the figures.

4.5.3 Cross Validation

5 Conclusão

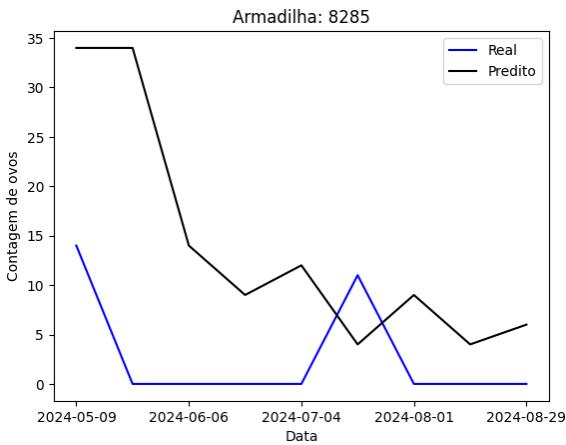
6 Apêndice

6.1 Tratamento de Valores Inconsistentes

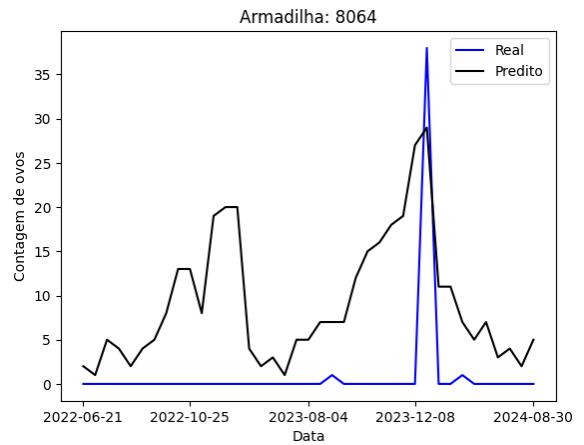
valores extremos foram tratados individualmente e, caso não possível, incorporados ao dataset como parte das imprecisões referentes ao processo de coleta, dado que representam porcentagem ínfima do dataset

6.2 Modelos iniciais

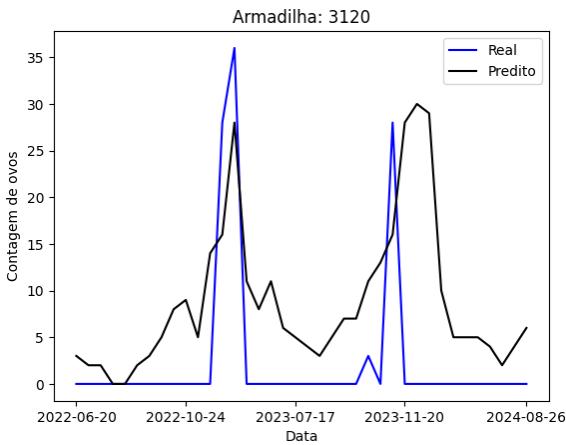
1: lags, days, lat long 68.9 2: semepid 68.9 (sem estrutura) 3: perc_{zero}69.05 : semedpid, semepid2, sin_{semepid}, (lat, long)daarmadilha70.06 : 100truncado70.01 : step70.0'logistic'sexterno70.0'logistic'cexterno - 70.3



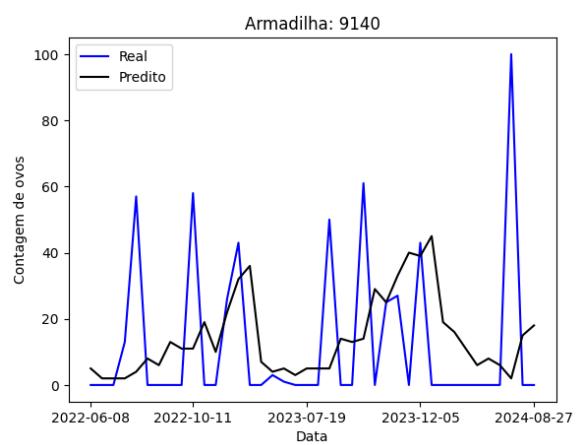
(a) First figure description.



(b) Second figure description.



(c) First figure description.



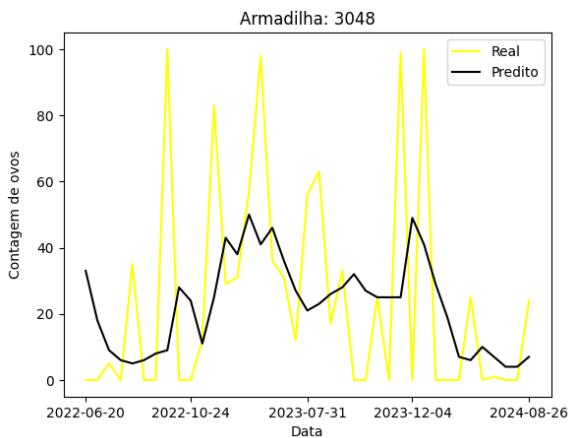
(d) Second figure description.

Figura 38: Overall caption for the figures.

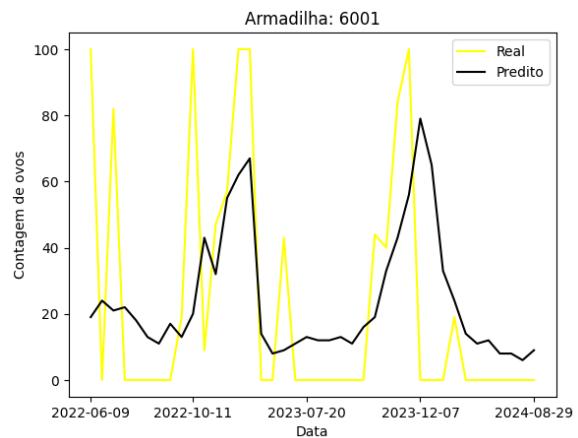
7 Bibliografia

Referências

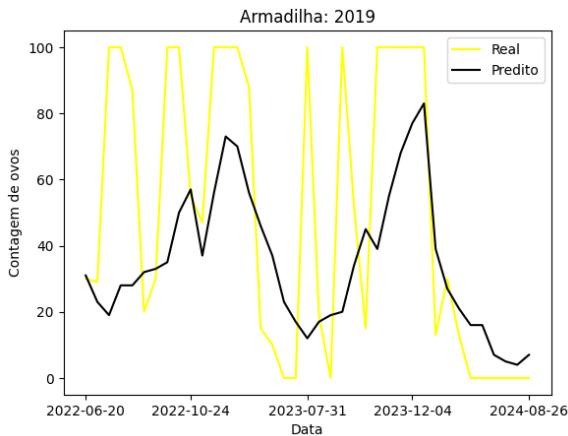
- [1] Thaddeus M Carvajal, Katherine M Viacrusis, Lara Fides T Hernandez, Howell T Ho, Divina M Amalin, and Kozo Watanabe. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan manila, philippines. *BMC infectious diseases*, 18:1–15, 2018.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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 sustai-



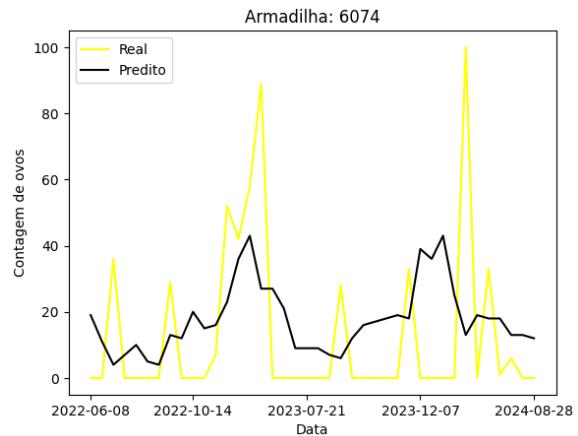
(a) First figure description.



(b) Second figure description.



(c) First figure description.

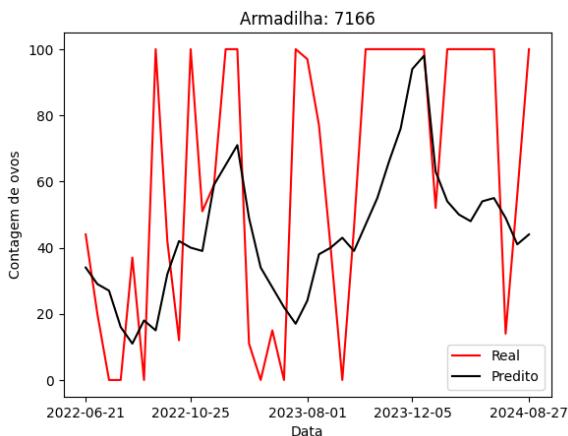


(d) Second figure description.

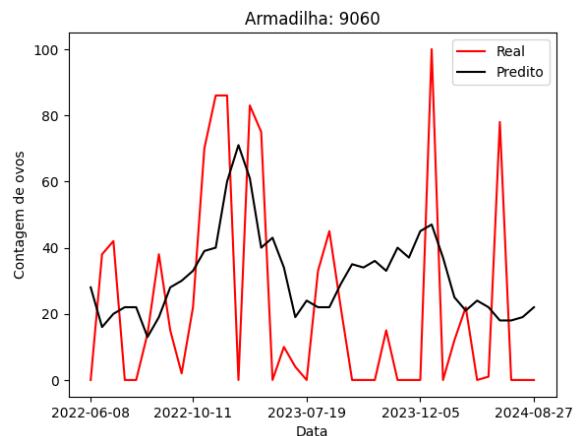
Figura 39: Overall caption for the figures.

nable health management in kharagpur city, india. *Spatial Information Research*, 27(2):187–204, 2019.

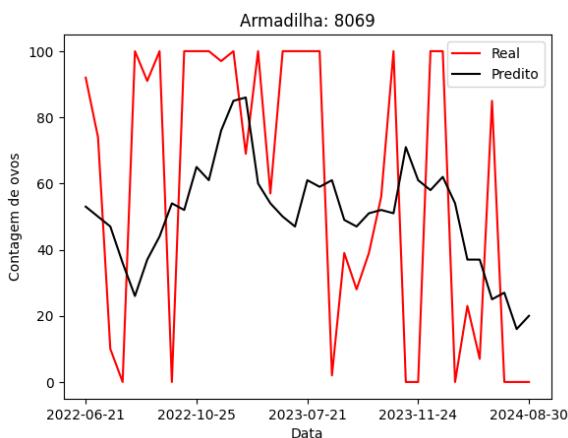
- [6] Instituto Brasileiro de Geografia e Estatística. Mapa climático do brasil - 2002. https://geoftp.ibge.gov.br/informacoes_ambientais/climatologia/mapas/brasil/Map_BR_clima_2002.pdf, 2002. Acesso em: 08 jun. 2024.
- [7] Instituto Brasileiro de Geografia e Estatística. Panorama de Belo Horizonte, MG. <https://cidades.ibge.gov.br/brasil/mg/belo-horizonte/panorama>, 2023. Acesso em: 07 jun. 2024.
- [8] Instituto Nacional de Meteorologia. Mapas meteorológicos do brasil. <https://mapas.inmet.gov.br/>. Acesso em: 10 jun. 2024.
- [9] Instituto Nacional de Meteorologia. Normais climatológicas do brasil. <https://portal.inmet.gov.br/normais>. Acesso em: 10 jun. 2024.
- [10] 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.
- [11] Benjapuk Jongmuenwai, Sudajai Lowanichchai, and Saisunee Jabjone. Comparision using data mining algorithm techniques for predicting of dengue fever data in northeastern of thailand. In *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 532–535. IEEE, 2018.



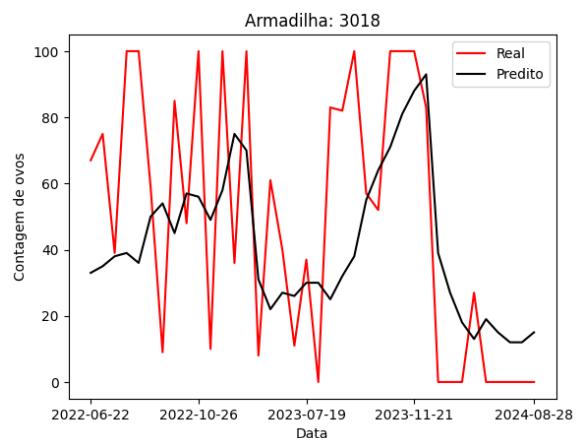
(a) First figure description.



(b) Second figure description.



(c) First figure description.



(d) Second figure description.

Figura 40: Overall caption for the figures.

- [12] N Kerdprasop, K Kerdprasop, and P Chuaybamroong. Computational intelligence and statistical learning performances on predicting dengue incidence using remote sensing data. *Adv Sci Technol Eng Syst J*, 5:344–50, 2020.
- [13] 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.
- [14] Cláisse 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.
- [15] 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.
- [16] 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.
- [17] Elsa Maria Nhantumbo, José Eduardo Marques Pessanha, and Fernando Augusto Proietti. Title of the article. *Revista Médica de Minas Gerais*, 22(3):265–273, Jul/Set 2012.

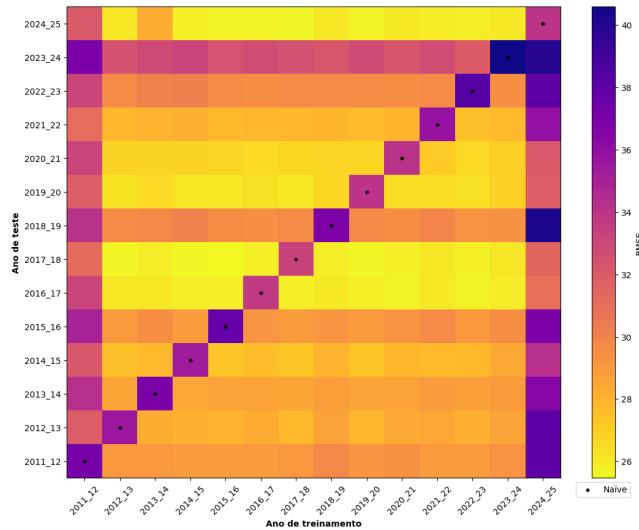


Figura 41: Caption

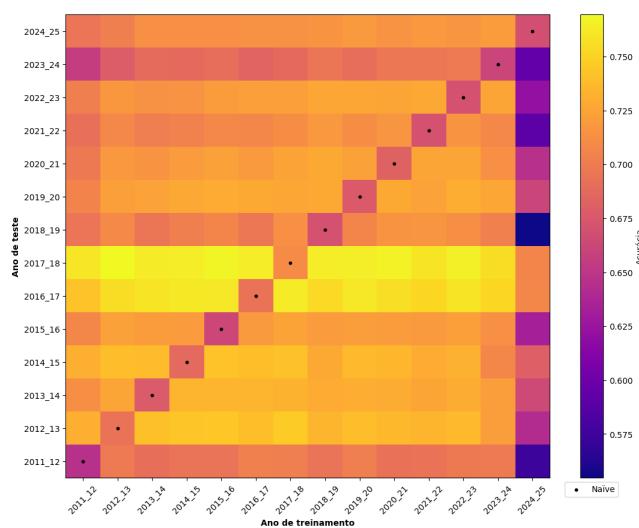


Figura 42: Caption

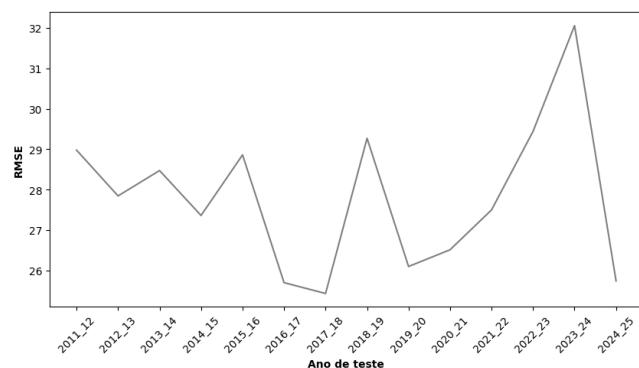


Figura 43: Caption

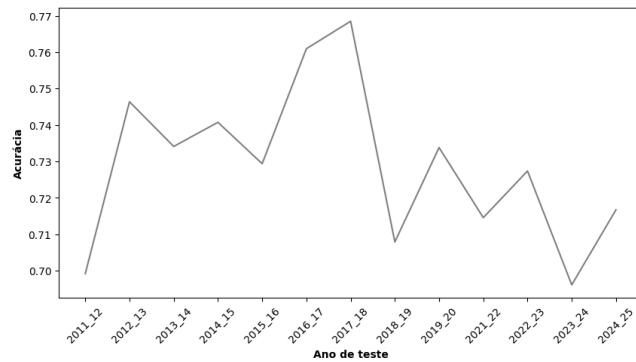


Figura 44: Caption

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

- [18] 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.

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

- [19] José Eduardo Marques Pessanha. Onde está wally? ou onde se esconde o aedes aegypti. *Boletim Epidemiológico*, X(4):26, 2007.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Sujit Sahu. *Bayesian modeling of spatio-temporal data with R*. Chapman and Hall/CRC, 2022.
- [24] 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.
- [25] 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.

- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Sediyama GC, Vianello RL, Pessanha JEM. Previsão de ocorrência dos mosquitos da dengue em belo horizonte com base em dados meteorológicos. 2006.
- [32] 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.