



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	Estado da Arte	5
3	Descrição dos Dados	6
4	Metodologia	19
4.1	Estrutura temporal e espacial	19
4.2	Variáveis Meteorológicas	21
4.3	Matriz de Entrada	23
4.4	Processamento e Divisão dos Dados	25
4.5	Escolha de Entradas	25
4.6	Modelos	26
4.7	Métricas	26
5	Resultados	27
5.1	Seleção das Entradas e do Processamento	27
5.2	Resultados dos Modelos Preditivos	33
6	Discussão	33
6.1	Catboost para Classificação de Presença	34
6.2	Catboost para Regressão	36
7	Conclusão	39
8	Bibliografia	45

1 Introdução

Infecções arbovirais, como Dengue, Chikungunya e Zika, transmitidas principalmente 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. (32) 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.(43)

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. Em especial, busca-se explorar a capacidade preditiva presente nos dados por meio da criação de modelos preditivos diversos.

O trabalho é iniciado, na Seção 2, com uma revisão bibliográfica, na qual se busca descrever o impacto das doenças arbovirais para o Brasil e a cidade de Belo Horizonte e as ações públicas para sua contenção, além de contextualizar o leitor quanto aos trabalhos realizados no sentido de descrever e modelar a dinâmicas da doença e de seu vetor. Em seguida, na Seção 3, Descrição dos Dados, o banco de dados de ovitrampas disponibilizado pela prefeitura é descrito e analisado. Dá-se especial enfoque às características espaciais e temporais das contagens de ovos. Já na Seção 4, a Metodologia, expõem-se os métodos utilizados para processamento desses dados, para a incorporação de variáveis meteorológicas e os modelos preditivos escolhidos para os experimentos. Os resultados necessários para o treinamento desses modelos e os resultados dos modelos em si são encontrados na Seção 5, Resultados. Tal seção é seguida pela Seção 6, Discussão, na qual os resíduos dos melhores modelos foram analisados na busca por padrões em seus erros. Por fim, na Seção 7, discorre-se sobre as principais conclusões do trabalho, além de possíveis alterações no escopo do projeto e na metodologia utilizada.

2 Pesquisa Bibliográfica

Arboviroses são doenças causadas por arbovírus (ARthropod BOrne VIRUS), um grupo de que inclui os vírus da dengue, Zika, chikungunya e Febre Amarela. ?? Dentre as arboviroses, a Dengue se destaca como a mais relevante em termos de impacto humano. ?? Estima-se que mais de 390 milhões de infecções por dengue ocorram anualmente, sendo 96 milhões sintomáticas, com graus de gravidade variáveis. ?? Além disso, a febre da dengue é a principal causa de mortes humanas entre as doenças transmitidas por vetores.?? O Brasil, em particular, ocupa a posição de país com o maior número de casos de dengue no mundo, de modo que essas doenças representam uma séria preocupação para sua saúde pública ??.

Há duas espécies principais de mosquitos responsáveis pela transmissão das arboviroses: *Aedes aegypti* e *Aedes albopictus*. Embora ambos estejam presentes em ambientes urbanos brasileiros, o *Aedes albopictus* demonstra preferência por ambientes rurais e silvestres. Por outro lado, o *Aedes*

aegypti apresenta características comportamentais que favorecem sua dispersão e adaptação em áreas urbanas, sendo raramente encontrado em áreas com pouca presença humana. Desse modo, há prevalência dessa espécie na transmissão das arboviroses no Brasil.

Seu ciclo de vida dura em média 30 dias, desde a postura do ovo, que eclodem de dois a três dias, passando por fase larval (5 a 7 dias) e de pupa (1 a 3 dias) até chegar finalmente à fase adulta que dura por volta de 20 dias, podendo aumentar para até 35 dias, a depender da temperatura, pluviosidade e altitude.

O comportamento de oviposição, em específico, desempenha um papel crucial na vida dos mosquitos, estando diretamente relacionado à sua sobrevivência e à transmissão das doenças. As fêmeas grávidas de *Aedes aegypti* são altamente seletivas na escolha dos locais para oviposição, demonstrando preferência por depósitos artificiais escuros, localizados em áreas sombreadas e com água limpa. Esses locais proporcionam condições ideais para a reprodução e o desenvolvimento das larvas, além de favorecerem a alimentação. Geralmente, tais ambientes estão próximos ao domicílio humano, o que facilita o ciclo reprodutivo do mosquito e aumenta sua interação com as populações. Entretanto, mesmo que não haja as condições ideais para postura dos ovos, eles são resistentes à falta de água e a baixas temperaturas, permanecendo viáveis por até 492 dias em períodos de seca. Uma vez imersos, os ovos eclodem rapidamente, iniciando o ciclo de vida.

Nesse sentido, o combate às epidemias de arbovírus concentra-se no controle do seu vetor por meio da eliminação dos seus criadouros. Visando reduzir o número de casos de febre amarela, já na metade do século XX, o combate ao *Aedes aegypti* no Brasil foi posto em prática. O controle vetorial era feito por meio da eliminação dos criadouros e aplicação de larvicidas, na tentativa de romper a cadeia de transmissão das doenças. Entre 1958 e 1973, o mosquito foi erradicado do país. Entretanto, em 1976, houve sua reintrodução, devido a falhas na vigilância epidemiológica e ao crescimento populacional urbano acentuado. Desde então a população do mosquito não foi suprimida. A partir de 1996, o Ministério da Saúde implementou o Plano de Erradicação do *Aedes aegypti* (PEAa), com o principal objetivo de reduzir os casos de dengue hemorrágica. Contudo, em 2001, o governo abandonou a meta de erradicar o mosquito e passou a focar no controle do vetor. Em 2002, a responsabilidade pela gestão e execução das ações de controle foi transferida para as secretarias municipais, com apoio do Ministério da Saúde e dos estados, permitindo ações de controle mais ajustadas às especificidades locais.

Tais ações de controle podem ser separadas em três tipos de abordagem preponderantes. As estratégias mecânicas visam eliminar os criadouros e reduzir o contato do mosquito com o homem, por meio da destruição ou descarte de locais de postura de ovos, drenagem de reservatórios e instalação de telas em portas e janelas. Nesse sentido, ações educativas junto à população visam garantir a eliminação contínua dos focos. Em segundo lugar, as estratégias de controle químico utilizam larvicidas e outros produtos para matar o inseto adulto ou em estágio larval. Por fim, abordagens biológicas são baseadas na utilização de predadores ou patógenos com potencial para reduzir a população do inseto, como peixes que se alimentam das larvas e pupas ou a bactéria Wolbachia, capaz de reduzir o tempo de vida de um mosquito adulto e tornar suas proles estéreis.

Complementando tais ações diretas, o monitoramento da população do vetor permite o mapeamento das regiões com alta atividade e o direcionamento de esforços às áreas mais afetadas, de modo a tornar as intervenções mais eficazes. Nesse sentido, uma das estratégias adotadas é a utilização de armadilhas, chamadas ovitrampas, para a coleta dos ovos do *Aedes aegypti*. Essa abordagem se baseia na forte correlação entre a contagem de ovos e a densidade da população de fêmeas do mosquito, responsáveis pela transmissão das doenças ao ser humano.

As ovitrampas, Figura (32), 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 período de exposição,



Figura 1: Exemplo de ovitrampa utilizada em Belo Horizonte - MG (32)

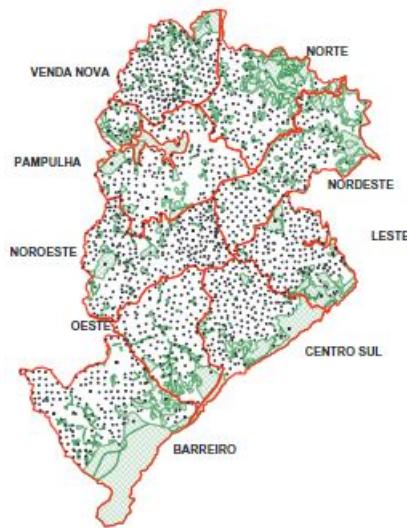


Figura 2: Grade de ovitrampas em Belo Horizonte (32)

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. Tais dispositivos são sensíveis à detecção das populações de *Aedes aegypti*, mesmo em períodos de seca, quando as populações estão reduzidas.

Em Belo Horizonte, em resposta à epidemia de dengue ocorrida no final de 1997 até maio de 1998, o controle do vetor da doença foi intensificado no contexto do PEAa. Como parte dos esforços de contenção às arboviroses, coletas sistemáticas de ovitrampas foram introduzidas em 2002. A partir de então, cerca de 1800 armadilhas espalhadas pelo perímetro urbano da cidade em locais fixos de uma grade regular (Figura 2), distando 200 metros entre si, são monitoradas. Nelas, placas são instaladas quinzenalmente em um padrão alternado, com instalação em quatro regiões em semanas ímpares e nas regiões restantes em semanas pares, e coletadas após sete dias de instalação. Esse padrão ocorre ao longo de todo o ano, com exceção de duas semanas no final do ano e durante o Carnaval. Após coletadas, as placas são levadas para o Laboratório de Entomologia da Prefeitura, onde os ovos são classificados e contabilizados. O número de ovos em cada armadilha varia comumente entre 10 e 50, podendo ultrapassar a ordem de milhar em locais de infestação severa. 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 (33).

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 do verão foi confirmado, assim como a sensibilidade

das ovitrampas em estações secas, períodos nos quais outros indicadores, como focos larvários, praticamente não são encontrados. (33)

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 (47) quanto embasam novos modelos. Por exemplo, (30) 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 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 de correlação de Pearson (ρ) consideráveis entre a média do número de ovos nas ovitrampas espalhadas pela cidade e o percentual de ovitrampas positivas ($\rho = 0.96$, p-valor < 0.01), também chamado de Índice de Positividade de Ovitrampa (IPO). Além disso, foram encontradas alta correlação da média dos ovos com a densidade de ovos em ovitrampas positivas ($\rho = 0.96$, p-valor < 0.01), a temperatura mensal ($\rho = 0.65$, p-valor < 0.01) e a precipitação mensal ($\rho = 0.54$, p-valor < 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 ($\rho = 0.72$, p-valor < 0.01). Desse modo, a viabilidade do uso 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 meteorológicos na previsão daquela. (32)

2.1 Estado da Arte

Em um contexto mais amplo, vários estudos foram 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. (27) 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. (27)

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.(46; 10; 50) Porém, parâmetros relacionados a variáveis obtidas por sensoriamento remoto(25), dados de redes sociais (45), consultas em ferramentas de busca (29) e, principalmente, dados relacionados ao monitoramento do vetor (11; 23; 28) também são ocasionalmente utilizados quando disponíveis.

Em relação às técnicas, os modelos de Poisson (20) e modelos de média móvel (ARIMA, SARIMA, ARIMAX) (12; 38; 34) são amplamente utilizados para previsão de séries históricas. Entretanto, constata-se o crescente uso de Redes Neurais Artificiais(39; 24), principalmente Long Short Term Memory (LSTM) (19), Máquinas de Vetores de Suporte (SVM) (42) e modelos baseados em Árvores de Decisão (41).

Retornando ao contexto da análise ovitrampas, novos estudos buscam aplicar modelos modernos para aperfeiçoar previsões. Modelos de aprendizado profundo foram utilizados em (40) 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. (43)

3 Descrição dos Dados

Os dados referentes à malha de ovitrampas de Belo Horizonte foram obtidas por meio da submissão de um projeto à prefeitura da cidade, conforme tutorial disponível no site da entidade (35), protocolado no dia 21 de junho de 2024.

Após a análise do projeto, foram disponibilizados dois grupos de dados. O primeiro grupo consistia em dados brutos exportados diretamente do sistema da Empresa de Informática e Informação do Município de Belo Horizonte (Prodebel), no qual as leituras das coletas são digitalizadas pelos técnicos da instituição. Essa base incluía informações de 2.066 armadilhas, com amostras coletadas entre setembro de 2011 e agosto de 2024, totalizando 524.387 registros. Cada linha da base representava uma coleta, contendo dados como a quantidade de ovos íntegros, eclodidos e secos aferidos, além das datas de instalação e coleta da placa e o endereço da armadilha, totalizando 32 colunas.

O segundo conjunto de dados, por sua vez, resultou do processamento e das análises realizadas pela equipe técnica da Prefeitura na base descrita anteriormente. Um total de 1.339 amostras, consideradas irrecuperáveis, foi excluído, e novas colunas, relacionadas às manipulações e categorização realizadas, foram acrescentadas. Dessa nova base, 5.249 amostras foram descartadas devido à ausência de informações sobre a quantidade de ovos. Embora outros valores faltantes ainda estivessem presentes, estes aparentemente não comprometiam a integridade das amostras, sendo mantidos para tratamento em etapas posteriores.

Em adição aos dois conjuntos descritos, também foram disponibilizados: um dicionário contendo a descrição de cada coluna da primeira base, que será incluído no Apêndice ??; as rotinas de tratamento utilizadas para a criação da segunda base; arquivos com as coordenadas de cada armadilha; e mapas das malhas. Informações adicionais relacionadas à segunda base, explicações sobre as análises realizadas e esclarecimentos sobre valores inconsistentes foram obtidos por meio de diálogo direto com representante da PBH.

Na Tabela 1 são apresentadas as colunas do segundo conjunto de dados selecionadas para este trabalho, acompanhadas de suas respectivas descrições. Ressalta-se que as colunas de ano, mês e semana referem-se ao ano epidemiológico, que difere do calendário comum ao iniciar-se em junho, conforme o padrão estabelecido pelos técnicos da prefeitura. Essa escolha visa alinhar a divisão anual com o ciclo sazonal do número de ovos, cujo pico ocorre nos meses de verão. Dessa forma, as contagens relacionadas a períodos de alta dentro do mesmo ciclo não são fragmentadas. A convenção de nomenclatura adotada para o ano epidemiológico utiliza o ano inicial do ciclo seguido

pelos dois últimos dígitos do ano subsequente, separados por um *underscore*. Por exemplo, no ano epidemiológico 2016_17, são consideradas as placas instaladas entre junho de 2016 e julho de 2017. Além disso, a coluna 'GerCat' refere-se às categorias geradas pela prefeitura para classificar as armadilhas de acordo com sua capacidade de proliferação. Cada armadilha é atribuída a uma de quatro classes de incidência de ovos: B (baixa), M (média), A2 (alta) e A1 (muito alta). Os detalhes dos cálculos para a atribuição dessas classes a cada armadilha encontram-se no Apêndice ??.

Colunas	Descrição
nplaca	Identificador da amostra
novos	Quantidade de ovos coletados na amostra
dtinstal	Data de instalação da armadilha
dtcol	Data de coleta da amostra
narmad	Identificador das armadilhas. Único 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 do local da armadilha
longitude	Longitude do local da armadilha
GerCat	Categoria da armadilha gerada pela Prefeitura

Tabela 1: Descrição das colunas utilizadas no conjunto de dados analisado

Dentre as colunas selecionadas, as referentes à localização das armadilhas apresentaram valores incorretos, como coordenadas ausentes ou incoerentes, além de registros com duas ou mais armadilhas atribuídas à mesma localidade. Após contato com a prefeitura, foi esclarecido que as armadilhas sem valor de latitude e longitude haviam sido desativas e armadilhas com coordenadas idênticas estavam instaladas em pontos distintos dentro de uma mesma área florestal. Diante disso, amostras com coordenadas inexistentes ou com valores absurdos foram descartadas. Já as armadilhas com a mesma localização foram mantidas, mas um pequeno valor foi adicionado às suas coordenadas para diferenciá-las durante a aplicação de métodos de agrupamento. No total, 49.316 amostras foram descartadas, restando 468.483 para análise, provenientes de 1.773 armadilhas distintas, cuja distribuição espacial está representada no mapa da Figura 3.

Dados incorretos também foram identificados nas colunas referentes às datas de instalação e coleta das placas. Os problemas incluíram datas de coleta anteriores à data de instalação, datas de coleta posteriores à data de entrega dos dados e intervalos anormalmente longos entre as datas de instalação e coleta. Após contato com a prefeitura, foi esclarecido que os erros nas datas provavelmente ocorreram durante a inserção dos dados na base e poderiam ser tratados e corrigidos. Assim, adotou-se um tratamento individualizado para cada amostra com valores inconsistentes, verificando-se que a maioria dos problemas estava relacionada à inserção incorreta das datas de coleta. Detalhes adicionais sobre o tratamento dos dados problemáticos estão descritos 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 concentraram consistentemente entre 6 e 8 dias, abrangendo 99% do total disponível. Outro valor avaliado foi a diferença entre as datas de coletas de placas da mesma armadilha, equivalente à sua taxa amostral (Figura 5). Apesar das adversidades que podem acarretar exclusão de amostras na base de dados, como seu descarte no processamento inicial, 80% das amostras apresentam intervalos entre 13 e 15 dias, com uma moda evidente em 14 dias.

Um segundo grupo relevante, equivalente a 8% do total, apresenta intervalos de aproximadamente 28 dias entre as coletas. Esse padrão pode ser explicado pelo fato de, consistentemente ao longo dos anos, uma amostra de cada armadilha não ser coletada no mês de dezembro e outra durante o período do Carnaval. Essa característica também é evidenciada na contagem de armadilhas por mês,

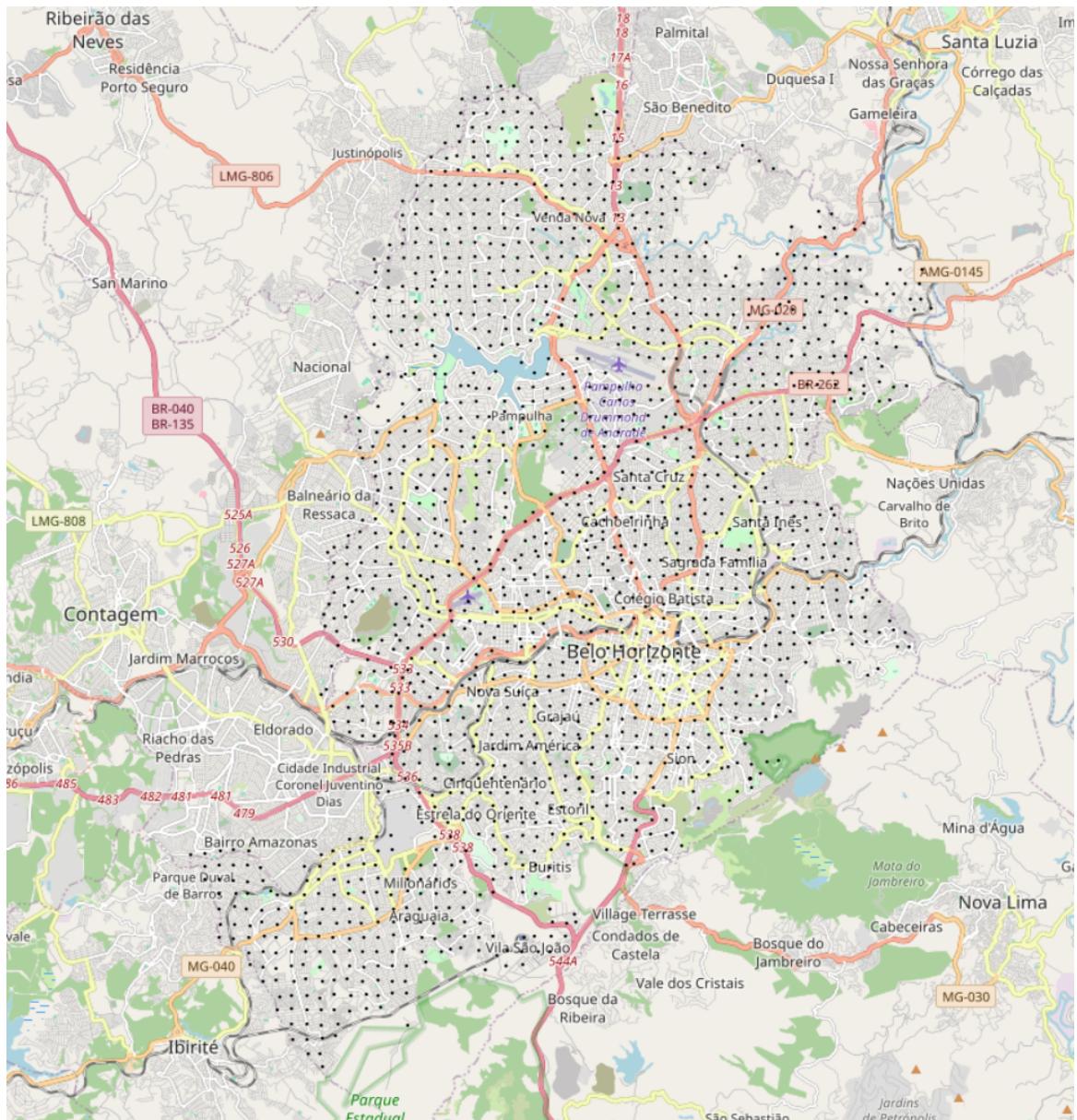


Figura 3: Mapa com a localização das armadilhas analisadas

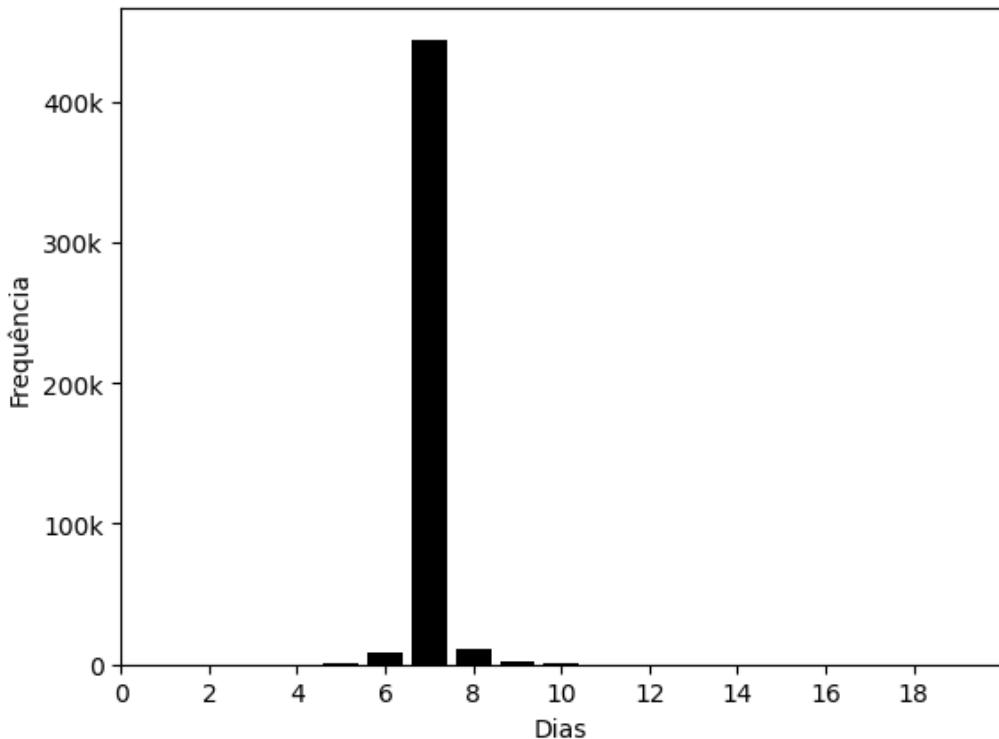


Figura 4: Histograma do tempo de exposição das placas

apresentada na Figura 6, onde o número de amostras nos meses epidemiológicos 7 (dezembro) e 9 (fevereiro) é aproximadamente a metade do observado nos demais meses.

Além da variação mensal, o número de placas também apresenta variação anual, referente ao processo gradual de implementação da política pública. Conforme Figura 7, o número de coletas demonstra tendência de crescimento anual, com exceção dos anos epidemiológicos de 2011_12 e 2024_25, por estarem incompletos na base utilizada. O mapa da Figura 8, por sua vez, ilustra o número de placas coletadas por armadilha. Nele, o círculo posicionado no local de cada armadilha tem raio proporcional à contagem de amostras 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 menor que a média das demais. Isso indica, em conformidade como exposto em conversas com o corpo técnico da prefeitura, a reestruturação da malha em anos recentes.

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 comparações. 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). Percebe-se também que 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 grandes epidemias (2012_13, 2015_16, 2018_19, 2022_23, 2023_24), (9). Pontua-se apenas a discrepância entre os valores médios dos anos de 2011_12 e 2024_25. Apesar do número de placas ser equivalente em ambos os anos, as coletas disponíveis do primeiro são referentes aos últimos meses epidemiológicos do ano, enquanto as amostras do último são referentes aos primeiros, respectivamente, períodos com maior e menor média associada.

Por outra perspectiva, a Tabela 2 apresenta o número de placas de cada categoria, a soma dos

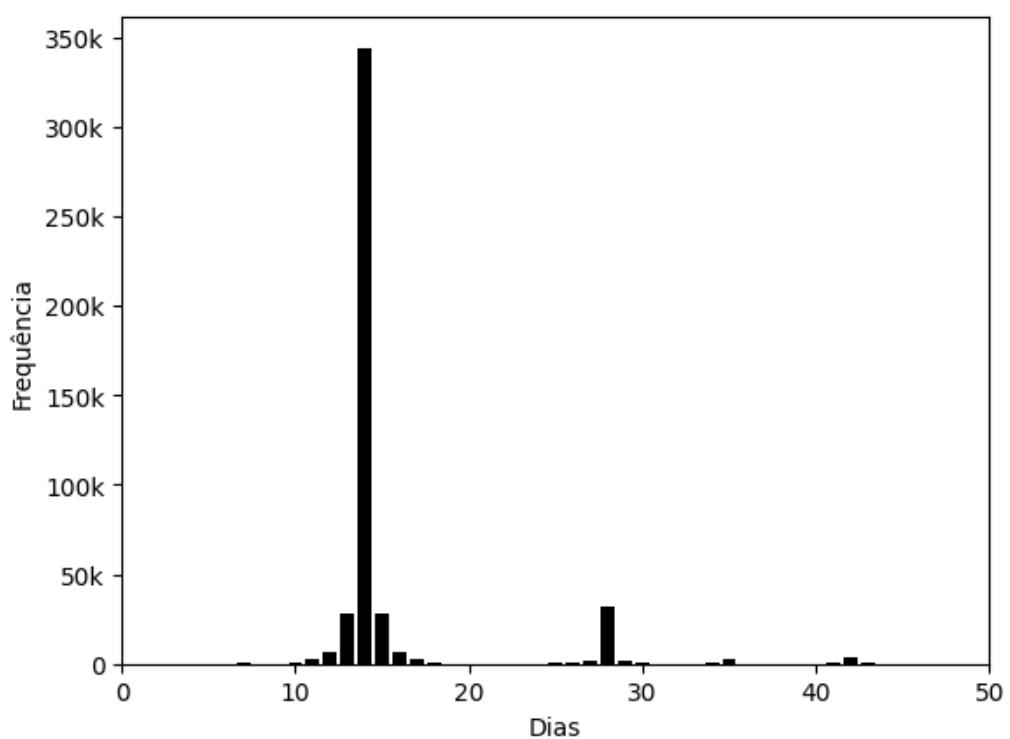


Figura 5: Histograma da frequência de amostragem das armadilhas

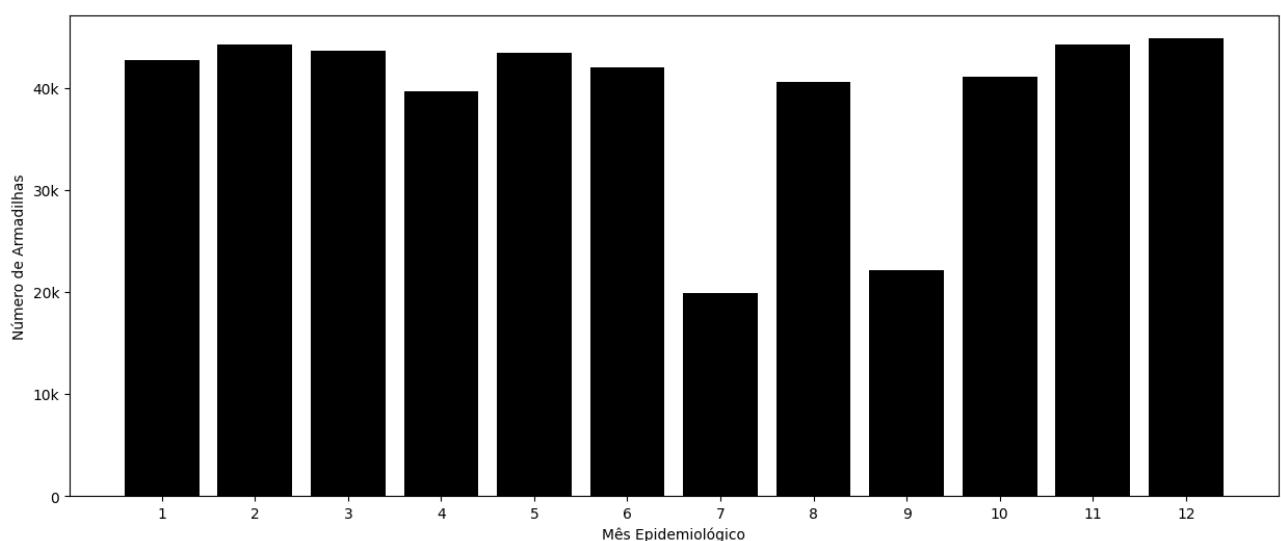


Figura 6: Número de placas por mês epidemiológico

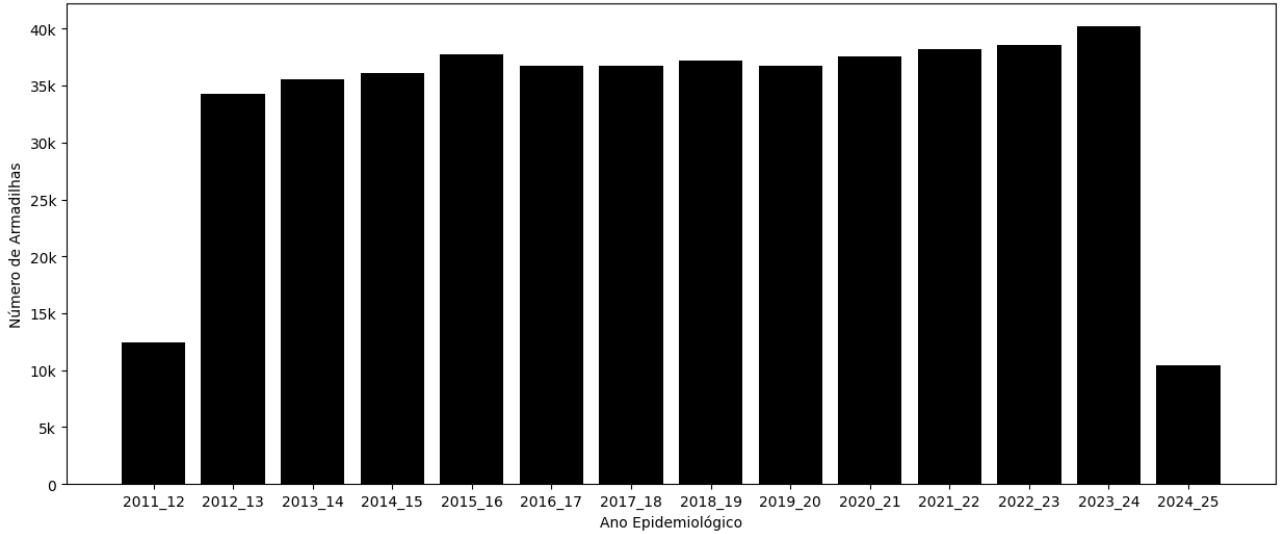


Figura 7: Número de placas por ano epidemiológico

ovos pertencentes a elas 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 média duas vezes maior que a da categoria A2. De maneira semelhante, armadilhas da categoria B compreendem apenas 3.0% do total de ovos, enquanto representam 12.8% das amostras.

Categoría	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 dessa heterogeneidade entre as categorias de armadilhas e iniciando uma análise com maior resolução do espacial em busca de identificar dinâmicas nesta dimensão, o mapa 11 foi gerado. A cada armadilha foi atribuída uma cor de acordo com a categoria à qual ela pertence 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. Agrupamentos de armadilhas dessas duas categorias podem ser identificados, sendo raras as aparições isoladas.

Com o objetivo de aumentar ainda mais a resolução dos estudos, as amostras individuais foram examinadas. Com uma média de 36 ovos por coleta, a contagem de ovos varia 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. Pela primeira imagem (12a), fica clara a prevalência de placas vazias e a alta divergência na quantidade 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.

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. Na imagem (12d), apenas amostras com mais de 1000 ovos são apresentadas, detalhando melhor esse comportamento. A partir desse valor, são frequentes contagens registradas apenas uma vez, que se tornam progressivamente mais esparsas à medida que os valores aumentam. Essa cauda pesada é característica de um grupo de distribuições, sendo a distribuição Pareto um dos representantes mais

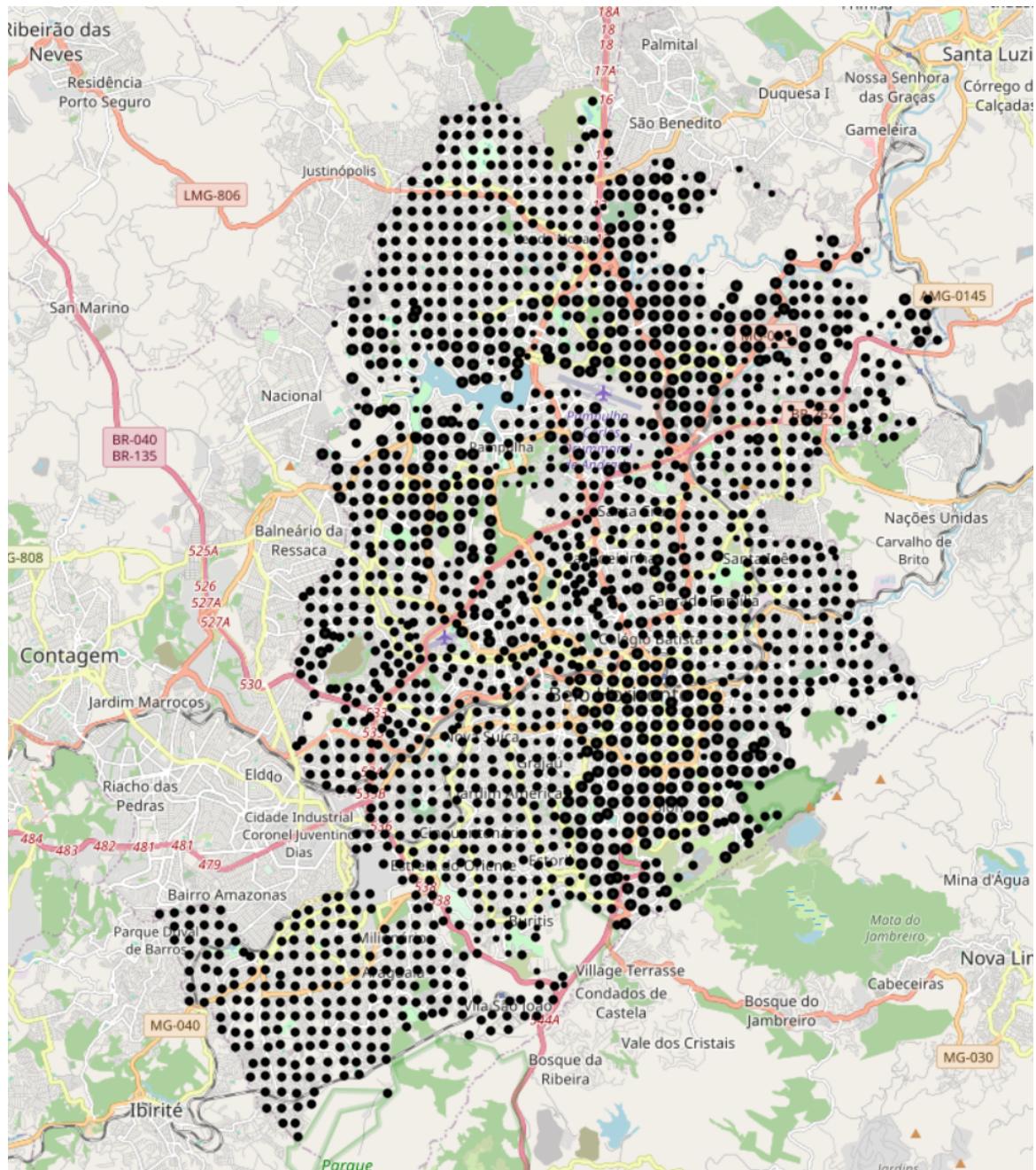


Figura 8: Número de placas por localidade

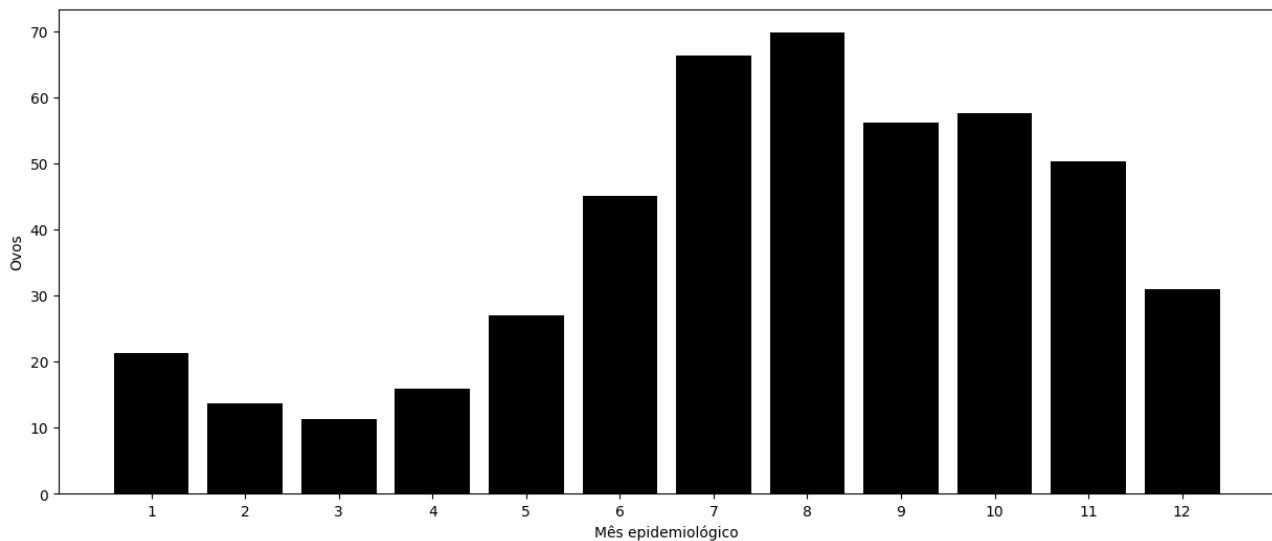


Figura 9: Média de ovos por mês epidemiológico

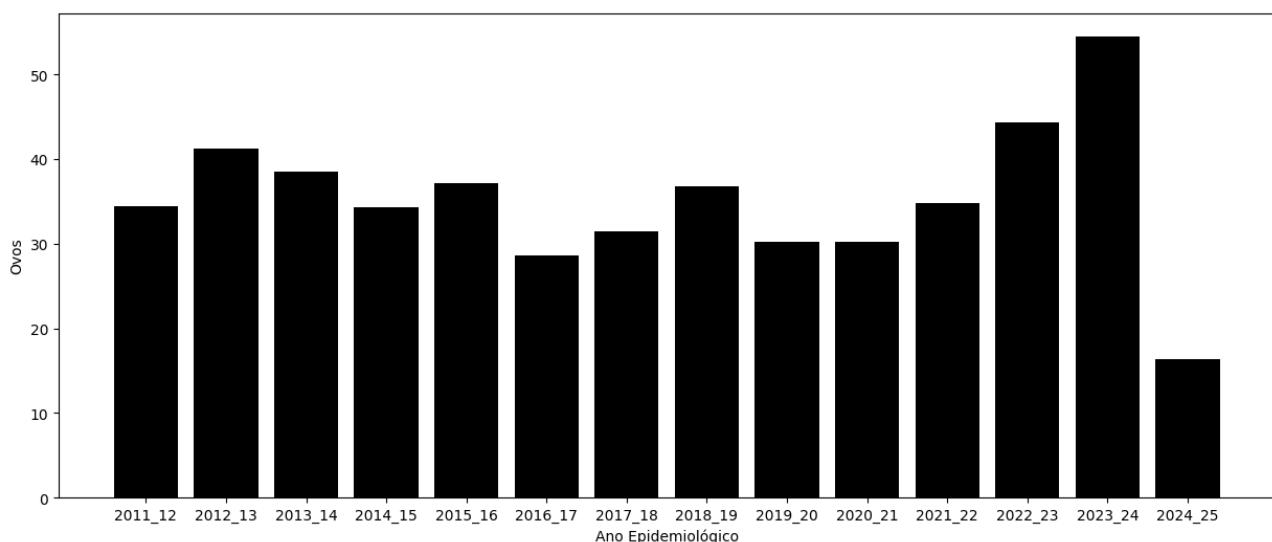


Figura 10: Média de ovos por ano epidemiológico

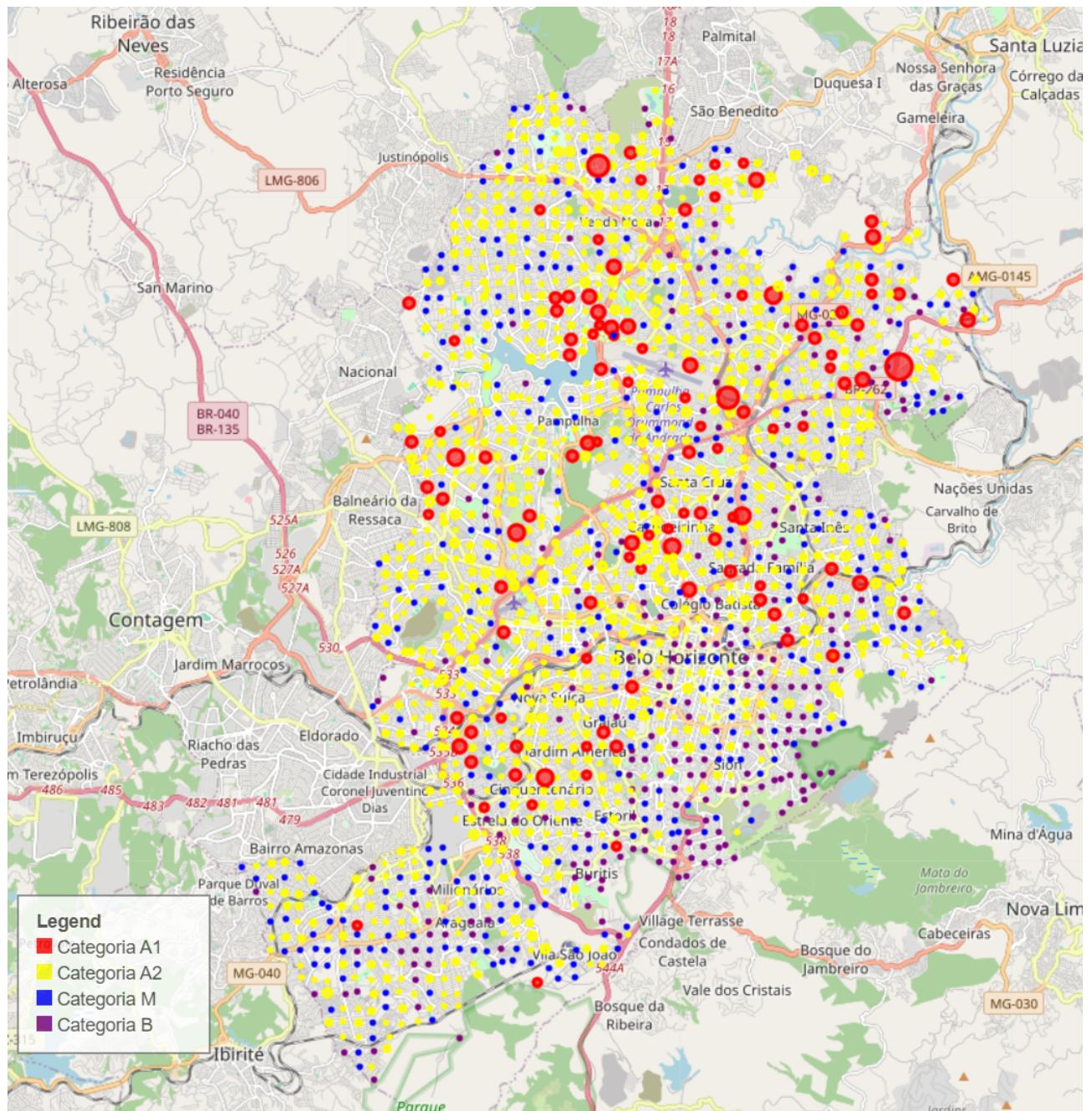


Figura 11: Mapa das armadilhas por classe

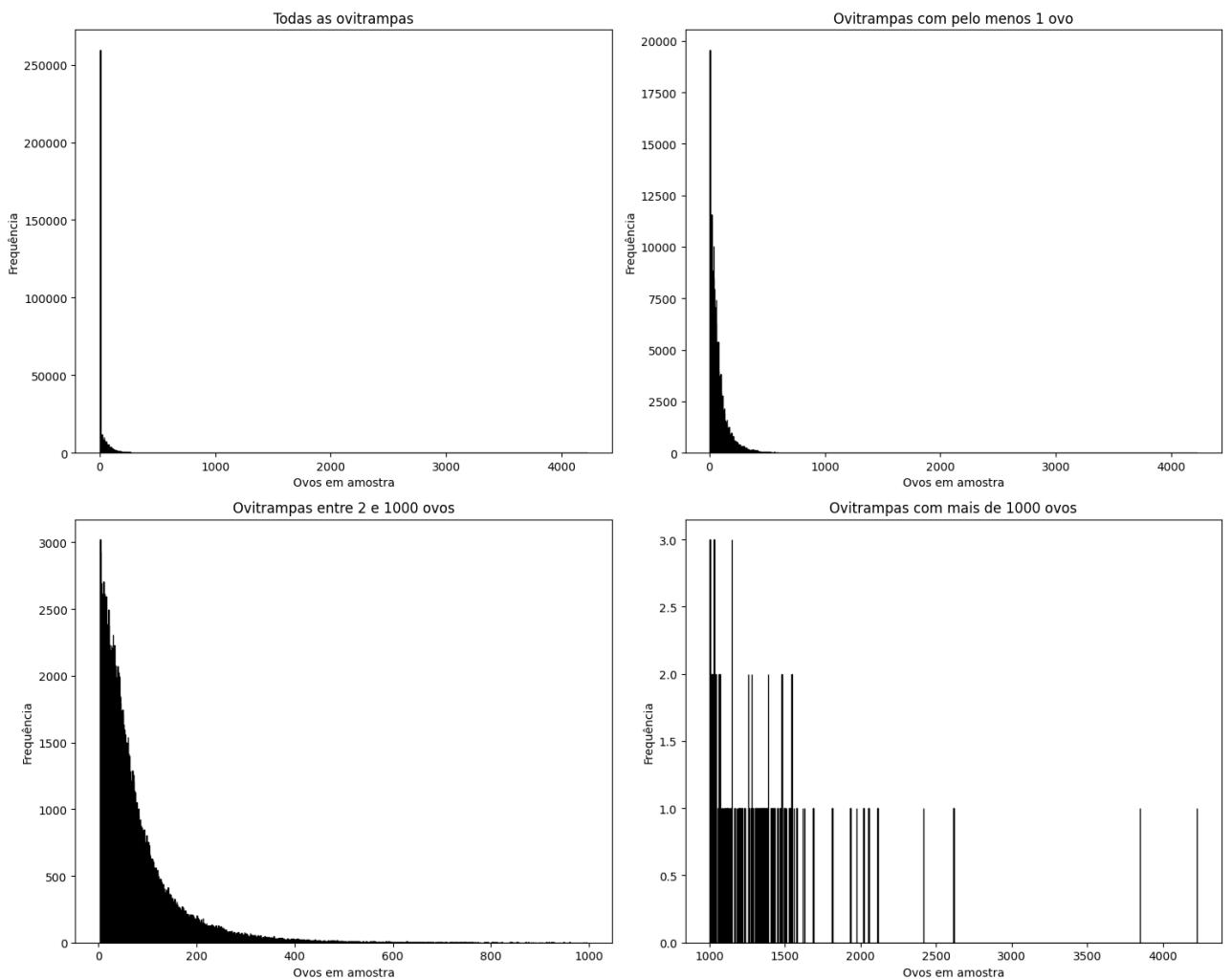


Figura 12: Histogramas das quantidades de ovos por ovitrampa. Da esquerda para a direita, de cima para baixo: (a) todas as ovitrampas; (b) ovitrampas com pelo menos um ovo; (c) ovitrampas entre 2 e 1000 ovos; (d) ovitrampas com mais de 1000 ovos

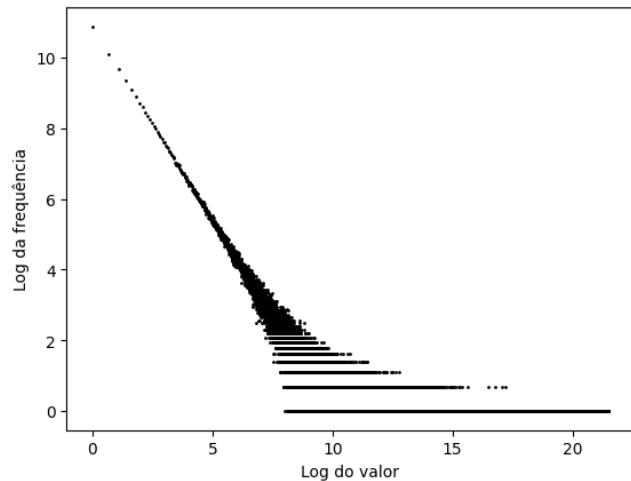


Figura 13: Gráfico log-log de uma distribuição Pareto gerada artificialmente

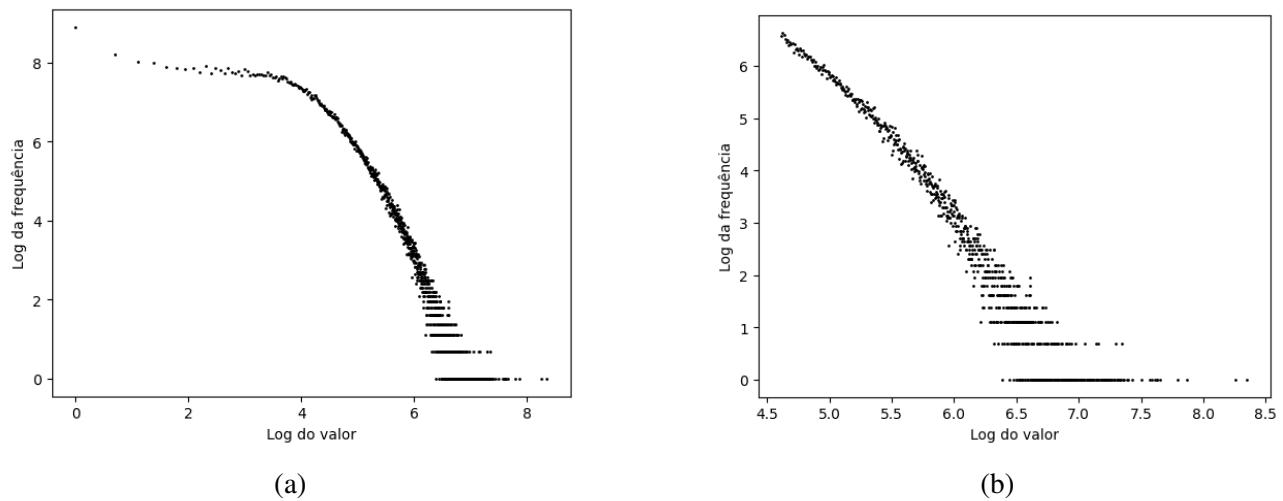


Figura 14: Gráfico log-log da distribuição de ovitrampas com todas as contagens de ovos são incluídas (a) e com apenas contagens a partir de 100 ovos (b)

conhecidos. Ela pode ser identificada por meio de um gráfico log-log, ou seja, pelo logaritmo da frequência em função do logaritmo do valor de interesse (3). A Figura 13, gerada a partir da função `zipf`, da biblioteca SciPy (48), 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 número, o gráfico 14a começa a se assemelhar ao 13, com uma relação aproximadamente linear seguida do espalhamento horizontal. Esse limite, portanto, será considerado o início aproximado da cauda pesada da distribuição de ovitrampas.

Outra característica de interesse da distribuição avaliada é a quantidade de amostras nulas, visto que pelo menos 50% das amostras totais estão vazias. O gráfico 15 apresenta a porcentagem de ovitrampas vazias do conjunto completo e separado por categoria. À direita, é apresentada a porcentagem das ovitrampas com valores não nulos. A relação entre a porcentagem de placas vazias e a média de ovos é clara. O gráfico 16, a seu turno, 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.

Por fim, para uma análise mais detalhada da estrutura temporal, as médias das amostras coletadas foram discriminadas por ano e mês epidemiológicos, como apresentado na Figura 17. De fato, constata-se componente mensal relevante, já que as tendências verificadas em um mês epidemiológico

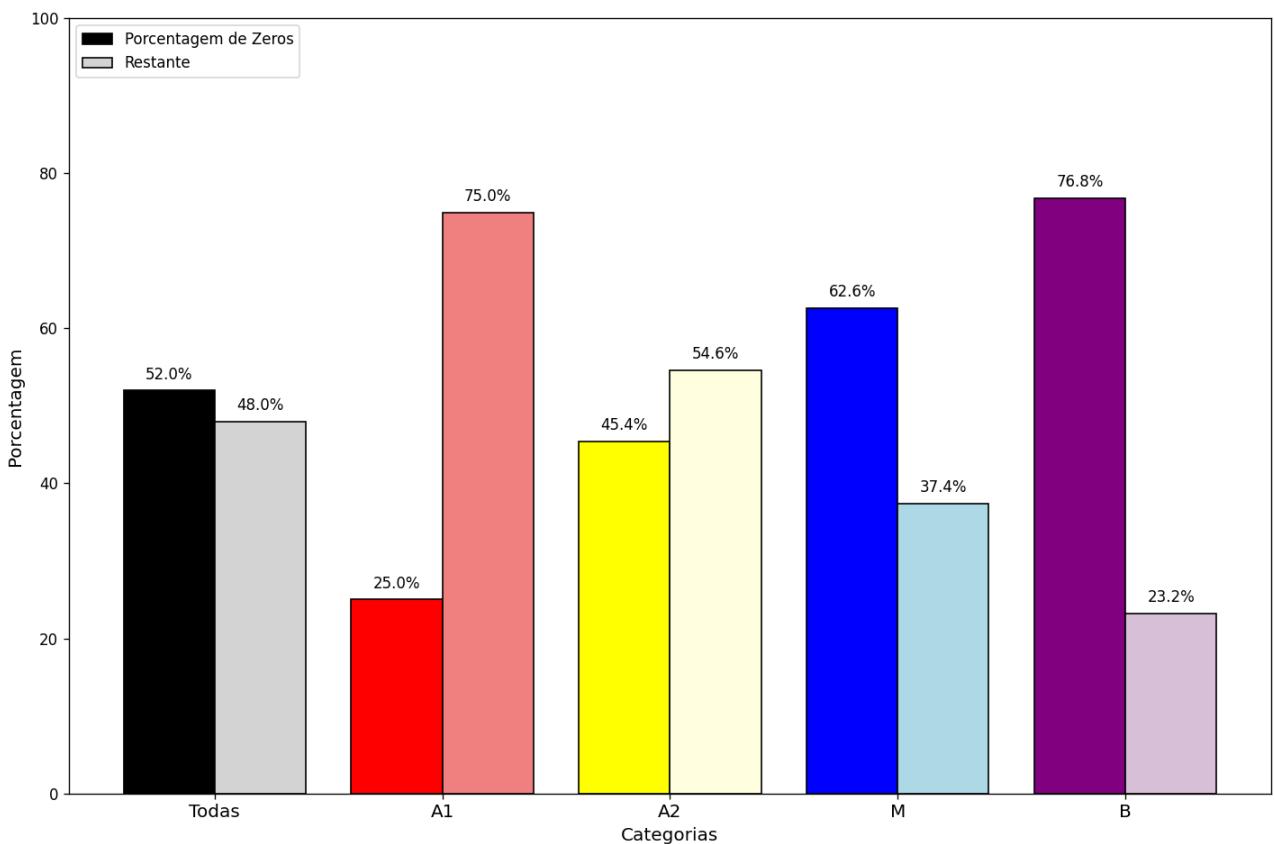


Figura 15: Porcentagem de ovitrampas sem ovos (à esquerda) e com ovos (à direita) em todo o banco de dados e por categoria

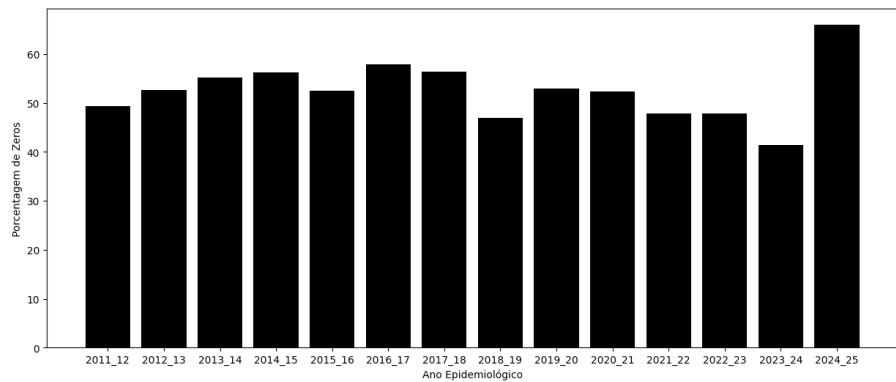
se repetem consistentemente ao longo do período. Essa análise, quando aplicada às quatro classes de armadilhas, corrobora a mesma conclusão, conforme ilustrado em 18.

Considerando as características expostas dos dados das ovitrampas e a proposta definida do projeto de avaliar a capacidade preditiva da base da PBH, três objetivos distintos foram definidos para os modelos explorados: (1) classificação da presença de ovos, (2) a previsão da quantidade de ovos e (3) a categorização de novas amostras em três faixas de valores, definidas em parceria com os técnicos prefeitura.

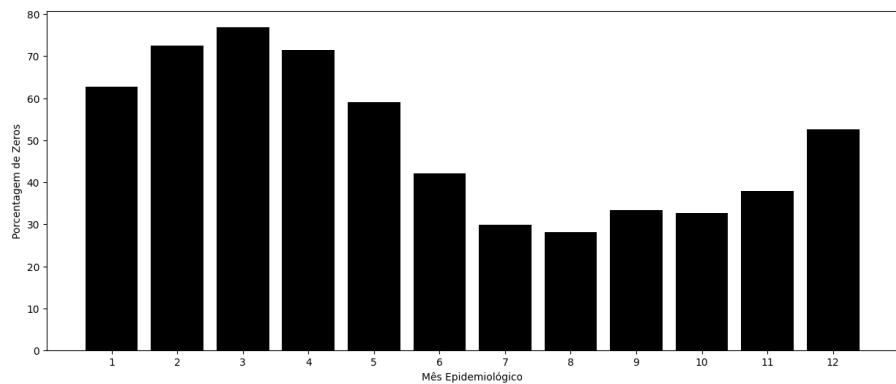
O problema (1) foi considerado o mais simples a ser trabalhado devido ao seu baixo número de classes ('presente' e 'ausente') e a seu balanceamento (Figura 15). Por isso, ele foi o enfoque inicial dos estudos e serviu como base para a tomada de parte das decisões do projeto. Ademais, caso bem-sucedido, o modelo desse problema poderia ser utilizado na solução dos demais. A previsão do número de ovos, por exemplo, poderia ter enfoque apenas em valores positivos, o que possibilitaria a aplicação de transformações na variável de saída.

O problema (2), por sua vez, caso concluído com sucesso, permitiria análises espaço-temporais mais detalhadas da dinâmica da postura de ovos. Entretanto, ele demanda um processamento à parte para lidar com as amostras com valores extremos de ovos, que certamente enviesariam modelos baseados na minimização da média dos erros. A proposta inicial de tratamento foi ajustar todos os valores superiores a 100, convertendo-os para exatamente 100. Esse valor foi definido a partir do limite aproximado da cauda pesada, descrita em 14

Finalmente, o problema (3) visa a aplicação das soluções pela prefeitura. Conforme conversas com o corpo técnico da instituição, parte dos estudos realizados limitam-se a verificar se a contagem das placas se encontra dentro de faixas de valores, sem grande preocupação com as variações internas a elas. Duas sugestões dos limites das faixas foram 2.65 e 5 ovos por dias de exposição. Considerando que menos de 0.01% das armadilhas mudaram de classe, tanto ao considerarmos os valores reais



(a)



(b)

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

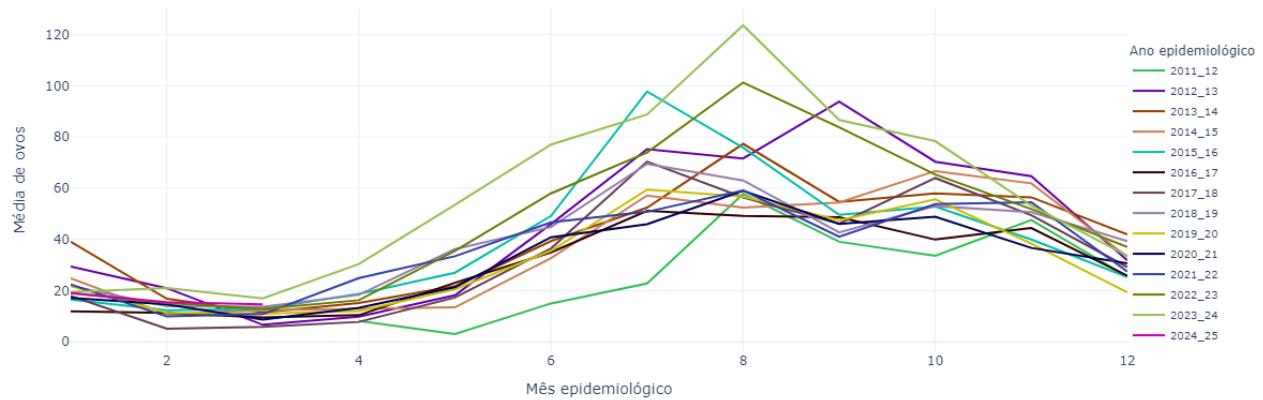


Figura 17: Série histórica da média de ovos separados por ano

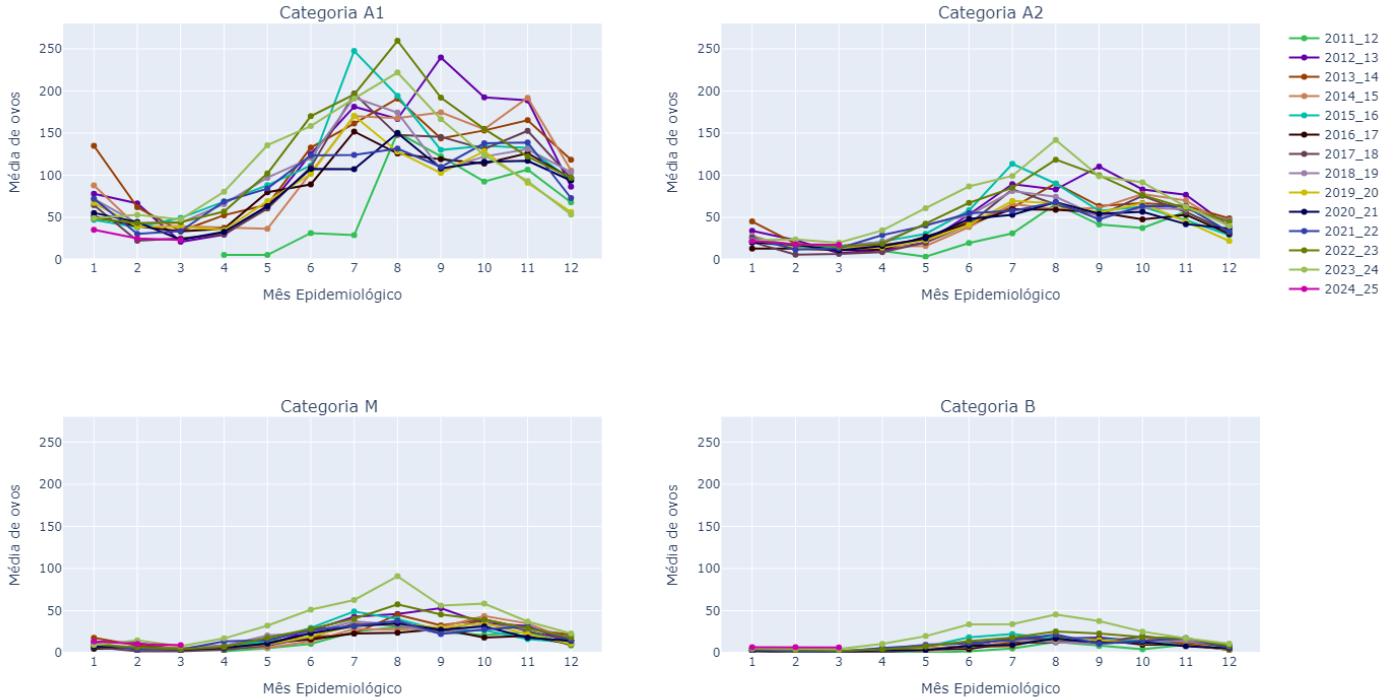


Figura 18: Série histórica da média de ovos separados por ano em cada classe de armadilhas

quanto ao definirmos como 7 o número de dias de exposição, padronizou-se o tempo de exposição para uma semana. Assim, amostras com menos de 19 ovos foram consideradas da classe 'baixa', amostras entre 19 e 35 ovos da classe 'média' e amostras com mais de 35 ovos, da classe 'alta'. Esta divisão adiciona um desbalanceamento considerável no grupo de dados, com 63.3% das amostras localizadas na primeira classe, 7.5% na segunda e 29.2% na terceira. Devido a esse desbalanceamento e ao fato de sua solução poder ser derivada da solução do problema de regressão (2), este problema foi, de forma geral, o menos explorado entre os três.

4 Metodologia

4.1 Estrutura temporal e espacial

Para introduzir uma estrutura espaço-temporal aos modelos testados, primeiramente, foram associados a cada amostra as coletas anteriores, tanto da armadilha em questão quanto de suas vizinhas. As quantidades máximas de armadilhas vizinhas e de valores passados de cada uma das armadilhas, tratadas respectivamente como t (*traps*) e l (*lags*), são dois dos hiperparâmetros do projeto. Por convenção, $t = 0$ se refere à própria armadilha e $l = 0$ à amostra analisada. Valores de t a partir de 0 e de l a partir de 1 foram avaliados. Assim, a quantidade imediatamente anterior de ovos na armadilha da coleta de interesse estará sempre inclusa nos modelos. O período entre as amostras atrasadas foi escolhido de modo que elas fossem coletadas na semana retrasada à placa analisada, mantendo a média de 15 dias entre as elas. As vizinhas, por sua vez, foram escolhidas com base na sua distância da armadilha principal e na disponibilidade de dados para o período analisado. Por exemplo, caso a armadilha mais próxima à principal tivesse suas coletas apenas a partir de 2020, em anos anteriores ela seria ignorada e outra armadilha seria considerada a mais próxima. Escolha semelhante não poderia ser feita para coletas cujos valores anteriores estivessem ausentes. Dada a variabilidade da contagem de ovos e a cauda pesada de sua distribuição, interpolar valores inexistentes

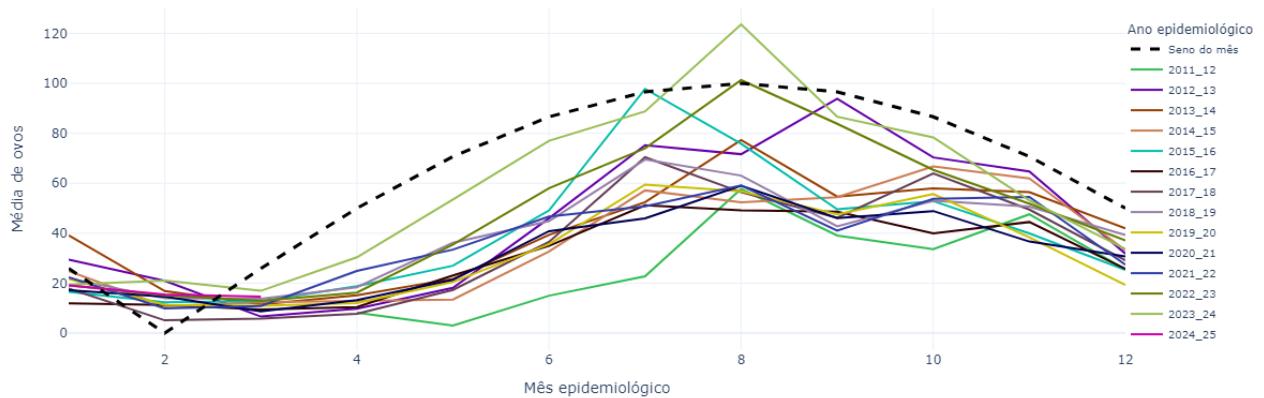


Figura 19: Comparação do seno do mês epidemiológico com a média mensal dos ovos segregadas anualmente

poderia enviesar os resultados obtidos, de sorte que tais amostras foram descartadas. Estes descarte foi propagado para demais contagens se uma das l amostras anteriores não tivesse valor válido. Portanto, aumentar l reduz a quantidade total de amostras disponível. Para o caso em que houvesse valores faltantes nas coletas anteriores das vizinhas, essa vizinha era desconsiderada conforme descrito e uma nova armadilha tomava seu lugar. O processo de descarte era repetido até que 50 armadilhas fossem descartadas, número arbitrário que visou reduzir o custo computacional do processo de criação do conjunto de amostras e evitar que armadilhas muito distantes fossem escolhidas.

As coordenadas centralizadas das armadilhas e de suas vizinhas e o mês epidemiológico da placa principal transformado em variável *dummy* foram escolhidos como o grupo de entradas principal dos modelos. Era esperado que modelos mais complexos tivessem a capacidade de inferir as respectivas estrutura espacial e temporal a partir delas. Ao mesmo tempo, modelos mais simples (o modelo linear para a regressão e o logístico para a classificação) foram treinados com essas variáveis e seu resultado foi utilizado como referência para avaliar o impacto das seguintes transformações:

- Substituição da variável categórica pelo seno do mês epidemiológico, alinhado como exposto na Figura 19
- Substituição da variável categórica pelo seno da semana epidemiológica, para a avaliação do impacto de uma resolução espacial maior
- Substituição da variável categórica pelos valores da semana epidemiológica e da semana epidemiológica ao quadrado

Em adição, um Modelo Aditivo Generalizado (GAM) foi implementado para testar a incorporação do spline da semanda epidemiológica como componente temporal e o produto tensorial das coordenadas das armadilhas como componente espacial, utilizando implementação da biblioteca pyGAM (44).

Elas foram feitas de sempre de forma individual e usadas para treinar os modelos citados.

Após definida a transformação com o melhor resultado, associou-se a cada amostra a porcentagem de placas nulas nas respectivas armadilhas no momento de instalação, a fim de considerar informações de um período de tempo mais longo. Essa foi uma alternativa às Categorias 'GerCat' da PBH, cujo processamento deveria ser replicado para cada amostra a fim de evitar vazamento de informação, o que foi considerado custoso para este projeto.

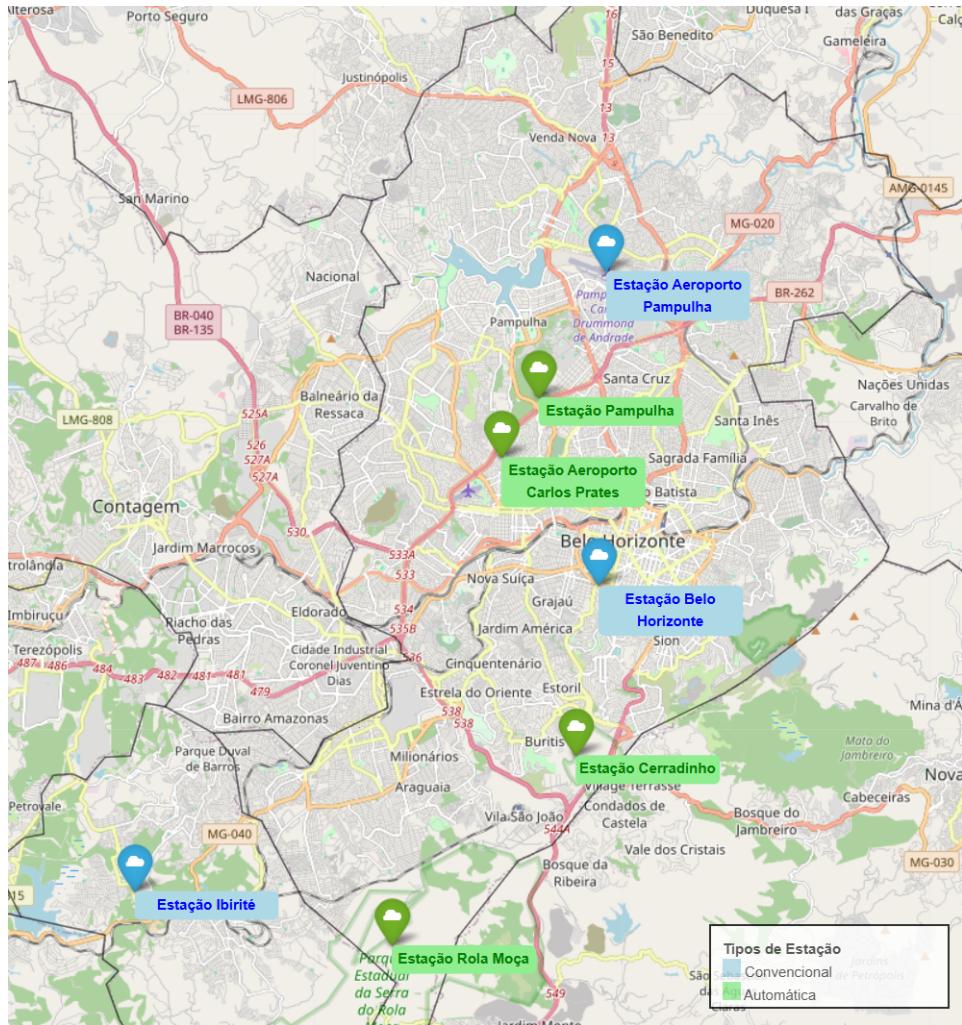


Figura 20: Localização das estações meteorológicas na região metropolitana de Belo Horizonte (22)

4.2 Variáveis Meteorológicas

Além dos dados das ovitrampas fornecidos pela PBH, no presente estudo foram utilizadas as variáveis exógenas temperatura, umidade e precipitação, visando adicionar novas informações espaciais e temporais às variáveis descritas na subseção anterior. Os dados utilizados tiveram como origem cinco estações meteorológicas do Instituto Nacional de Meteorologia (INMET) e duas estações do Departamento de Controle do Espaço Aéreo (DECEA), obtidos por meio do acesso à base de dados das próprias instituições (22), (17), (15), (16).

As estações cobrem a região metropolitana de Belo Horizonte em uma malha não regular, conforme mapa da Figura 20, coletando dados por períodos com início e finais distintos, em alguns casos não cobrindo o período completo de coleta das ovitrampas. Além disso, os dados dos pontos de coleta foram registrados com taxas amostrais variáveis em 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). Tendo em vista também a incompatibilidade espacial entre esta malha e a malha das armadilhas, foi necessário processamento prévio para conformação dos dados meteorológicos aos entomológicos. Para tal, seguiu-se a seguinte rotina de tratamento.

Inicialmente os valores diários das três variáveis climáticas foram agregados por meio de sua média, de modo que as frequências amostrais das sete estações fossem as mesmas. Em seguida, os valores das variáveis para cada um dos dias do período analisado foi associado às armadilhas por meio do inverso do quadrado das distâncias (IDW), conforme equação 1. Assim, associou-se à cada armadilha j 7 valores, proporcionais ao inverso do quadrado das distâncias às 7 estações, que, após

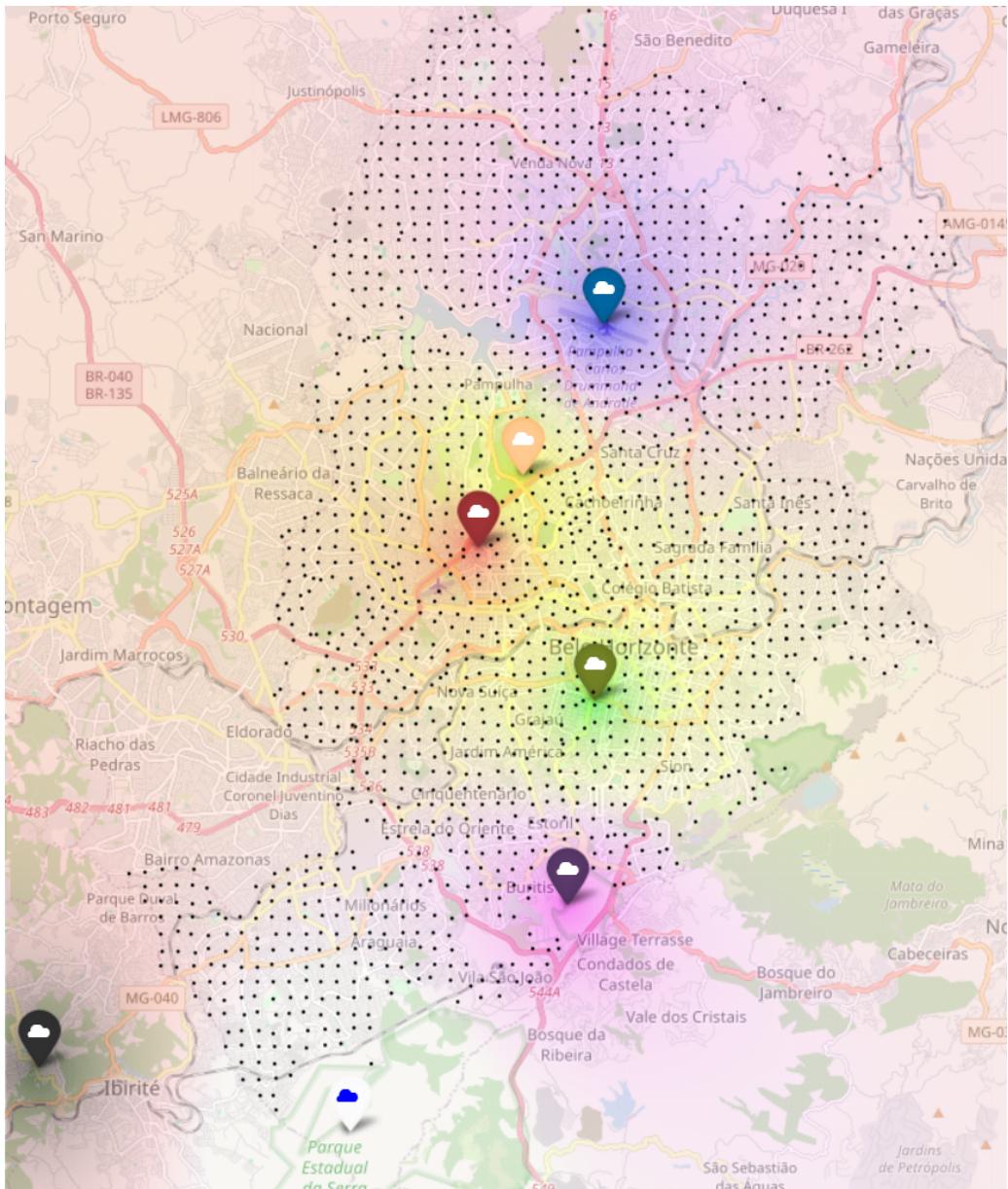


Figura 21: Influência de cada estação na malha de armadilhas, representada pela mistura de cores

normalizados, foram utilizados como pesos da média ponderada da variável meteorológica M. Caso não constasse valor de alguma das variáveis da estação i, no dia k, ela era ignorada no cálculo da média. O mapa 21 reproduz por meio de cores a influência das estações em cada armadilha. Cada estação foi representada por marcadores de cores distintas e a cor da região da armadilha é a mistura das cores de cada estação ponderadas pelos mesmos pesos calculados pela equação 1.

$$M_j^{<k>} = \frac{\sum_{i=1}^7 \frac{1}{d_{ij}^2} \cdot M_i^{<k>}}{\sum_{i=1}^7 \frac{1}{d_{ij}^2}} \quad (1)$$

Tendo em mãos as três matrizes relacionando cada variável exógena a uma armadilha nas datas de interesse, restava atribuir a cada coleta os valores de temperatura, umidade e precipitação. Para isso, com intuito de explorar diferentes cenários, foram escolhidos dois períodos, utilizados de modo independente. No primeiro, a média dos valores da semana anterior à data de instalação da placa foi associada a cada coleta, de modo a ser possível trabalhar com medições existentes. A segunda escolha foi utilizar o período de exposição da placa como referência, calculando novamente a média dos valores

nesses dias, no intuito de mapear as condições meteorológicas no momento da postura. Apesar de, no dia de instalação da placa, não haver disponibilidade desses dados exógenos, para aplicações dessa escolha em modelos reais é possível utilizar por previsões meteorológicas de entidades especializadas. Em conclusão, a cada placa foram associados seis valores, a Precipitação, Umidade e Temperatura médias da semana anterior à instalação e do período de exposição, referenciados no restante do trabalho pelos sufixos `_previsao` e `_week_bfr_mean`.

Globalmente, análises das três variáveis escolhidas confirmavam correlação exposta na literatura (33). As médias mensais de cada uma delas foram comparadas à média mensal da contagem de ovos, como exposto nos gráficos à esquerda da Figura 22. Em seguida, a correlação cruzada entre as séries foi mensurada, indicando correlação não desprezível.

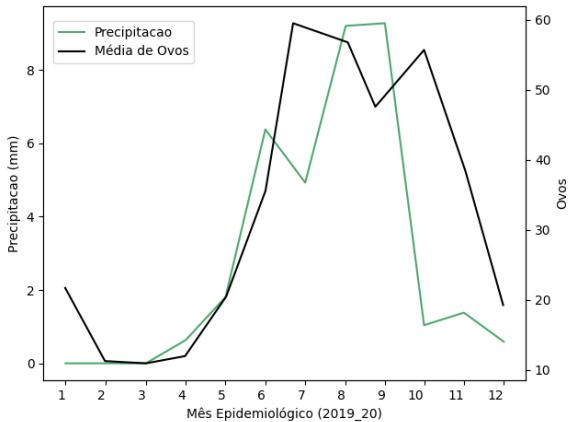
4.3 Matriz de Entrada

Após a adição das variáveis exógenas e das variáveis transformadas, a matriz de entrada continha as seguintes colunas:

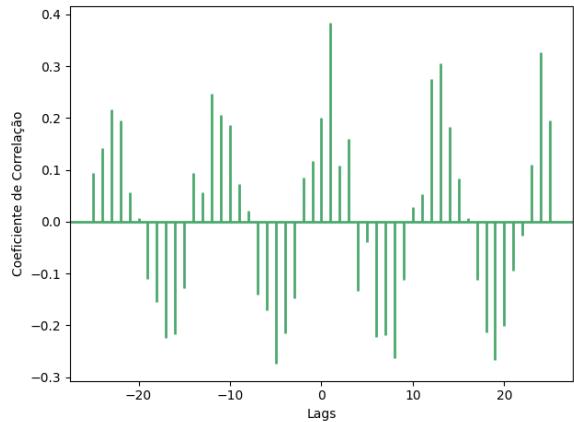
Coluna	Descrição
<code>nplaca</code>	Identificador da amostra
<code>narmad</code>	Identificador das armadilhas. Único por local de coleta
<code>anoepid</code>	Ano epidemiológico de instalação da amostra
<code>novos</code>	Quantidade de ovos coletados na amostra
<code>trap{t}_lag{1}</code>	Quantidade de ovos coletados na armadilha vizinha t com atraso de 1 amostras
<code>latitude{t}</code>	Latitude da armadilha vizinha t
<code>longitude{t}</code>	Longitude da armadilha vizinha t
<code>mesepid</code>	Mês epidemiológico de instalação da amostra, transformado em variável categórica <i>dummy</i>
<code>sin_mesepi</code>	Seno do mês epidemiológico de instalação da amostra
<code>semeipi</code>	Semana epidemiológica de instalação da amostra
<code>semeipi2</code>	Quadrado da semana epidemiológica de instalação da amostra
<code>sin_semeipi</code>	Seno da semana epidemiológica de instalação da amostra
<code>zero_perc</code>	Porcentagem de amostras com zero ovos para aquela armadilha desde o início da coleta
<code>Temperatura_previsao</code>	Temperatura média no local da armadilha nas datas de exposição
<code>Precipitacao_previsao</code>	Precipitação média no local da armadilha nas datas de exposição
<code>Umidade_previsao</code>	Umidade média no local da armadilha nas datas de exposição
<code>Temperatura_week_bfr_mean</code>	Temperatura média no local da armadilha na semana anterior à instalação da amostra
<code>Precipitacao_week_bfr_mean</code>	Precipitação média no local da armadilha na semana anterior à instalação da amostra
<code>Umidade_week_bfr_mean</code>	Umidade média no local da armadilha na semana anterior à instalação da amostra

Tabela 3: Nome e descrição das variáveis de entrada

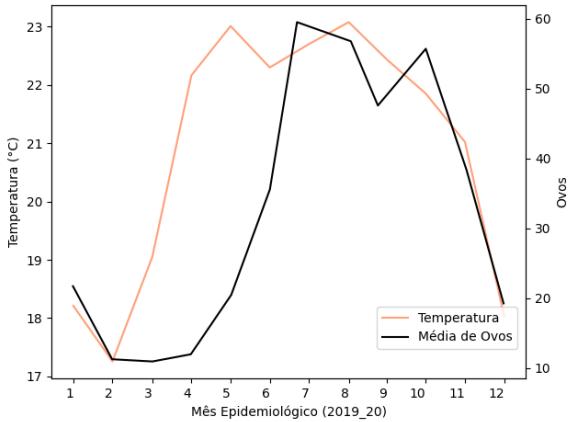
As variáveis `nplaca` e `narmad`, apesar de não serem entradas, foram mantidas para a identificação da placa.



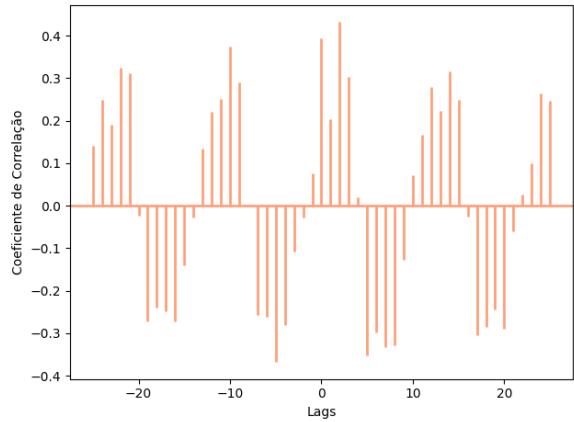
(a)



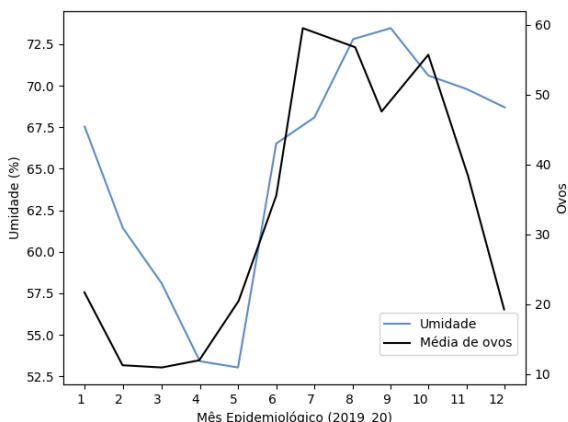
(b)



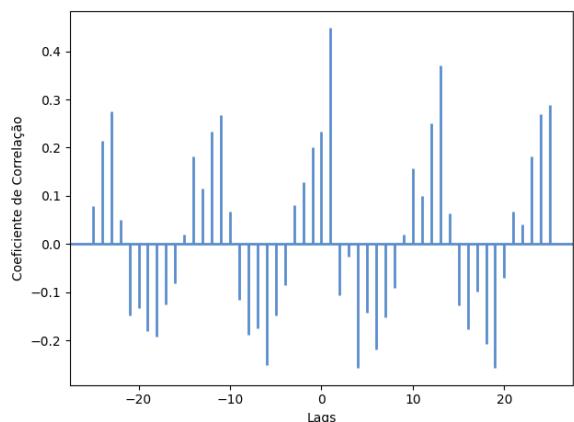
(c)



(d)



(e)



(f)

Figura 22: Séries das médias mensais das três variáveis meteorológicas sobrepostas à de ovos (direita) e correlação cruzada das séries diárias normalizadas de ambas as variáveis (esquerda)

4.4 Processamento e Divisão dos Dados

Os mesmos processamentos realizados na elaboração dos três problemas objetivos foram aplicados às variáveis de entrada para tentar amenizar o impacto das amostras extremos: o número de ovos foi truncado em 100, substituído por uma variável booleana indicando a presença ou ausência de ovos nas coletas, ou substituído por uma das três classes representando faixas de valores. Os modelos de referência foram testados com estes três processamentos e os resultados comparados aos modelos com entradas sem processamento.

Todos modelos, os de referência e os finais, foram treinados com os dados das coletas de 2011_12 a 2021_22 e testados com as amostra de 2022_23 a 2024_25, com o intuito de conservar a verossimilhança do problema e evitar vazamento de informações. O grupo de treino foi normalizado utilizando o normalizado MaxAbsScaler, da biblioteca Scikit-learn (31). Em seguida, o grupo de teste foi escalonado com os mesmos parâmetros.

4.5 Escolha de Entradas

Ao final da introdução das variáveis espaço-temporais e meteorológicas, o número de entradas do modelo seria da ordem de $(l + 1) \cdot (t + 2)$. Para o caso de 10 armadilhas escolhidas (a armadilha principal mais 9 vizinhas) com 10 amostras de atraso, 100 entradas seriam relativas às amostras atrasadas da armadilha e de suas vizinhas, 20 entradas referentes às coordenadas das armadilhas e por volta de 10 entradas relacionadas à estruturação espaço-temporal, incluindo as três meteorológicas. No sentido de aumentar interpretabilidade dos modelos criados e evitar instabilidades numéricas, as entradas foram filtradas. Outro ponto considerado foi o número de amostras disponíveis para valores diferentes de vizinhos e coletas atrasadas, já que, como descrito, o aumento do número de armadilhas e coletas passadas incluídas pode reduzir a quantidade de dados à disposição.

Para a escolha das entradas foi seguida a metodologia descrita por Aguirre (1). Entradas com baixa correlação com a variável de saída foram descartadas. Dentre as restantes, a correlação entre elas foi calculada e variáveis com alta correlação com outras entradas foram descartadas. Entretanto, a correlação de Pearson foi substituída pelo cálculo da Informação Mútua entre as variáveis, dado que esta é capaz de mensurar correlações não-lineares e não depende de assumpções de normalidade, como aquela (6), (14). Os limites escolhidos para manutenção e descarte das variáveis foram baseados na magnitude relativa dos coeficientes em comparação às demais entradas e à própria variável. Essa metodologia foi aplicado ao problema de regressão e ao de classificação da presença dos ovos, com resultados semelhantes. Todas as variáveis foram normalizadas para ambos os problemas e apenas as amostras do conjunto de treinamento foram consideradas para esse cálculo. Dado o caráter exploratório das análises deste trabalho, pouca atenção foi dada a estimativa prévia das densidades das variáveis, como sugerido em (13). Isso, no entanto, não alterou as conclusões do trabalho de acordo com os resultados complementares da seguinte metodologia.

Baseada no Critério de Informação de Akaike (AIC), métrica que valoriza o ajuste do modelo aos dados e penaliza a adição de novas variáveis (1), a segunda metodologia foi inspirada na Regressão Stepwise, implementada para linguagem R (37), e utilizado somente no problema de classificação para efeito de comparação. Diversos modelos logísticos foram treinados com grupos distintos de entradas e o grupo que resultasse no modelo com maior AIC era escolhido. Três políticas de mudança de variáveis foram adotadas, 'foward', 'backward' e 'bidirectional'. Na primeira, iniciava-se com um modelo vazio e a entrada que mais aumentasse o AIC era adicionada. Novas entradas eram adicionadas até que o conjunto de variáveis não utilizadas não resultasse em melhorias no modelo. A política 'backward' seguia o caminho inverso, começando com um modelo contendo todas as variáveis, que eram retiradas individualmente até que um critério de parada semelhante fosse atingido. Por fim, a abordagem 'bidirectional' combinava ambas as anteriores. Assim que a primeira política parasse de adicionar variáveis, a segunda era seguida para filtrar as entradas escolhidas. Após a parada da segunda política, o ciclo de adição e remoção era repetido até que o grupo de variáveis que iniciou o

ciclo fosse preservado.

4.6 Modelos

A escolha de modelos preditivos adequados impacta diretamente na capacidade de capturar as estruturas espaço-temporais da malha de ovitrampas. Assim, foi realizada uma análise de diversos modelos para avaliar o impacto das variações na complexidade, arquiteturas e métodos de treinamento e representação nos resultados, sendo os modelos selecionados com base na literatura apresentada e nas características dos dados. Os modelos mais simples escolhidos foram o modelo logístico para os dois problemas de classificação e o linear para o problema de regressão.

Dois modelos baseados em Árvores de Decisão foram escolhidos para aproveitar a estrutura temporal estratificada dos dados: o Random Forest (RF) (5) e o CatBoost (36), (21). Um Multi-Layer Perceptron (MLP) foi escolhido por sua capacidade como aproximador universal de funções, com o objetivo de explorar a possibilidade de mapear as coordenadas e a data de coleta para o número de ovos, sem a necessidade de tratamentos prévios mais complexos (7). Por fim, modelos baseados em Support Vector Machines (SVM) foram selecionados devido à sua capacidade de discriminar amostras essenciais para a separação de classes e regressão (7), (?).

Naturalmente, os modelos dessas famílias foram adaptados para os problemas estudados, com as devidas modificações para ajustá-los aos problemas de classificação, regressão e classificação de múltiplas classes. Quando aplicável, a escolha dos hiperparâmetros foi feita avaliando-se 20 grupos de valores amostrados aleatoriamente a partir de uma grade previamente definida, os quais foram utilizados para treinar os modelos. O grupo que resultou no melhor modelo conforme a métrica de avaliação foi separado. Esse processo foi repetido cinco vezes para cada modelo, e os grupos finais foram avaliados para refinamento. Idealmente, esse procedimento seria repetido em grades progressivamente menores, até que se atingisse a convergência em valores ótimos. Detalhes da implementação de cada modelo e da escolha de seus hiperparâmetros se encontram no Apêndice ?? Em todos os modelos de regressão, as saídas foram arredondados para o inteiro mais próximo e valores negativos, caso existissem, transformados em 0 a fim de garantir a validade das previsões;

4.7 Métricas

A fim de quantificar a diferença entre os valores reais e previstos, o desempenho dos modelos de regressão e classificação foi avaliado, respectivamente, por meio da Raiz do Erro Quadrático Médio (RMSE), medida em ovos, e da Acurácia, duas métricas amplamente utilizadas na literatura. (26)

O modelo ingênuo, que utiliza o valor atrasado da amostra como predição, foi empregado como referência. Esse modelo é comumente utilizado como ponto de partida em projetos preditivos, justamente por ser simples e intuitivo. Ele exige pouco processamento para ser aplicado e, frequentemente, apresenta bons resultados (18). No caso das malhas de ovitrampas, a coleta exatamente anterior à amostra de interesse, realizada na mesma armadilha, apresenta poder preditivo considerável para o problema de classificação de presença, como exposto na matriz de confusão 4. O simples ato de repetir o resultado da amostra anterior garante acurácia de 67.5% no conjunto de dados compreendendo todo o período disponível.

	Anterior: Ausente	Anterior: Presente
Atual: Ausente	37.5%	16.2%
Atual: Presente	16.0%	30.3%

Tabela 4: Matriz de Confusão Percentual do modelo Ingênuo para classificação da presença de ovos

De forma análoga, para o problema de regressão, a amostra anterior indicou ser uma boa variável preditora. No gráfico da Figura 23, busca-se ilustrar esse comportamento posicionando as amostras

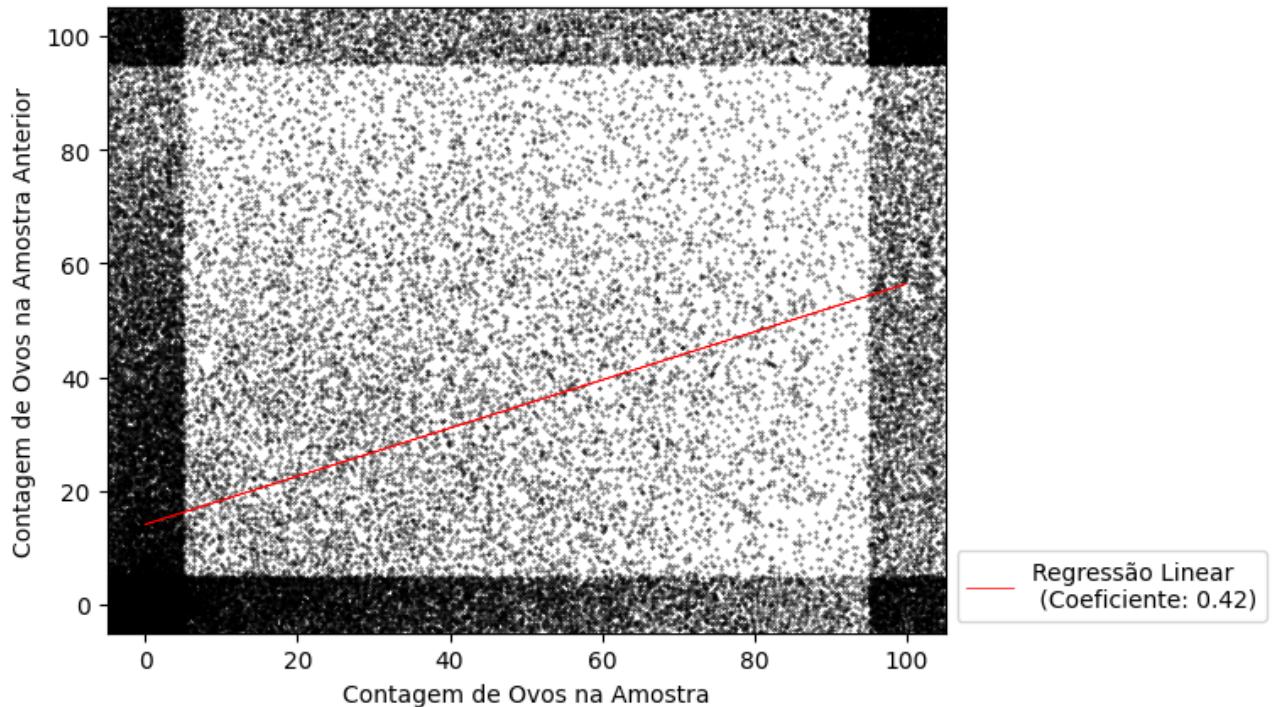


Figura 23: Número de ovos da amostra atual vs. Número de ovos da amostra anterior

com a contagem da coleta atual no eixo x e a contagem da coleta anterior no eixo y. O processamento dos dados é semelhante ao que foi proposto para lidar com as caudas pesadas das distribuições. Todos as amostras maiores que 100 foram truncados nesse valor. Foram adicionados a cada ponto um valor randômico entre 0 e 5 e grafados apenas 20% das amostras, escolhidas randomicamente, para melhorara a interpretação dos agrupamentos. A Regressão Linear em vermelho foi calculada considerando todas as amostras. No caso de um modelo ingênuo perfeito, seu coeficiente seria 1 e todas as amostras estariam na diagonal do gráfico. O coeficiente calculado demonstra correlação intermediária, assim como nuvem de pontos encontrada próxima à coordenada (20,20). Por outro lado, as duas nuvens que acompanham as bordas do gráfico adicionam uma nova faceta a essa interpretação. Valores pequenos, próximos a zero, apesar da tendência de serem acompanhados por contagens próximas, não raramente são antecedidos ou sucedidos por valores em toda a escala. Além disso, o truncamento em 100 concentrou os valores extremos nas regiões superior e direita do gráfico. Assim, o modelo ingênuo, apesar de exprimir parte do comportamento da variável, deixa espaço para a introdução de modelos mais complexos.

5 Resultados

5.1 Seleção das Entradas e do Processamento

O primeiro passo da Seleção de Entradas foi definir valores iniciais para os hiperparâmetros t e l , relativos ao número de vizinhos da placa e ao número de valores atrasados da armadilha (lags), dado que esta escolha altera diretamente o número de amostras disponíveis. Considerou-se razoável para análise inicial avaliar combinações de 1 a 20 armadilhas e 1 a 20 valores passados. Para cada um dos 400 pares de hiperparâmetros, um DataFrame foi gerado conforme descrito em na seção 4.1. A quantidade de amostras disponíveis em cada um desses DataFrames foi contabilizada para gerar o gráfico de superfície 24. Por meio dele, verifica-se que o número disponível de amostras depende praticamente apenas do número máximo de valores passados. Quando esse número é 1, por exemplo, aproximadamente 40 mil amostras são descartadas do conjunto original, restando 420 mil amostras.

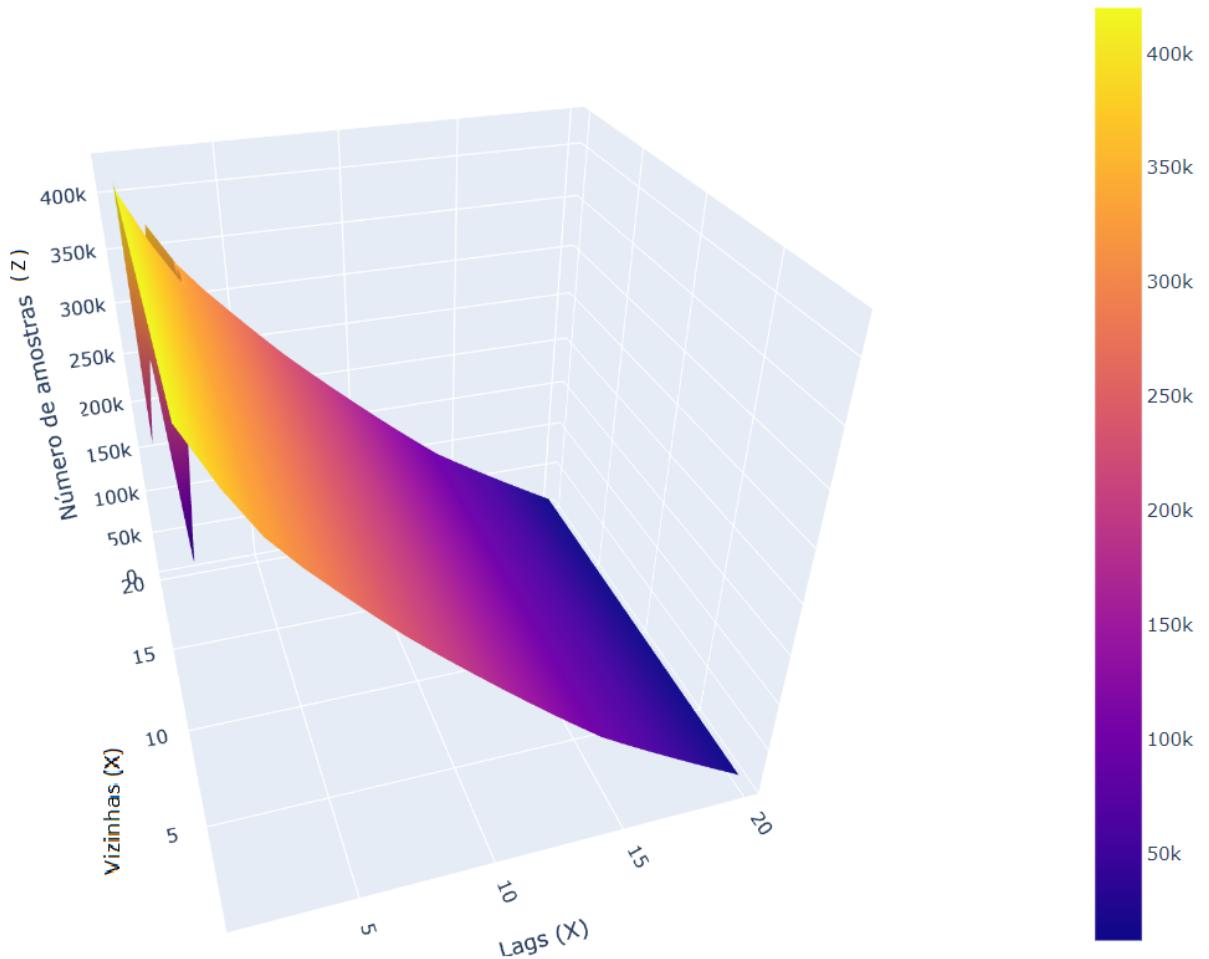


Figura 24: Número de amostras no conjunto de dados com diferentes números máximos de armadilhas vizinhas e valores atrasados

Para $l = 2$, restam 373 mil. Já para um número de lags máximo igual a 8, metade das amostras originais são desconsideradas. Assim, definiu-se que o número máximo de lags não seria maior que 14, para que a quantidade de amostras disponíveis não seja reduzida para menos que 100 mil. Não houve alteração no número máximo de vizinhos.

Com estes dois valores definidos, calculou-se a informação mútua entre o número de ovos e as variáveis de entrada (25). A entrada com maior correlação com a saída, como esperado, foi a contagem de ovos anterior, seguida pelas variáveis para a estruturação espaço-temporal e pelas coletas da própria armadilha ordenadas cronologicamente, grupo caracterizado por coeficiente de alta magnitude em comparação às demais variáveis. A partir do sexto lag, entretanto, os coeficientes não se destacam tanto e armadilhas vizinhas começam a aparecer como relevantes. Além disso, perde-se a estrutura temporal que ordenava as amostras da armadilha principal a partir do 7º atraso, em que a armadilha do 14º atraso apresenta coeficiente maior que a do 8º, por exemplo. Aliada a esta divisão, constatou-se a inexistência de amostras de seis meses epidemiológicos, consequência da escolha do atraso associada à ausência sistemática de coletas nos meses de dezembro e fevereiro. Por isso, optou-se por reduzir ainda mais o atraso máximo para 5 amostras, de modo que contagens de nove meses epidemiológicos estivessem presentes. O número de vizinhos também foi reduzido para 10 devido à falta de relevância das variáveis mais distantes.

Outra constatação baseada nesta primeira análise foi que apenas as coordenadas da armadilha principal demonstraram relevância. Além de ser intuitivo concluir que as armadilhas próximas terão coordenadas parecidas, a informação mútua da latitude e longitude da armadilha principal com as

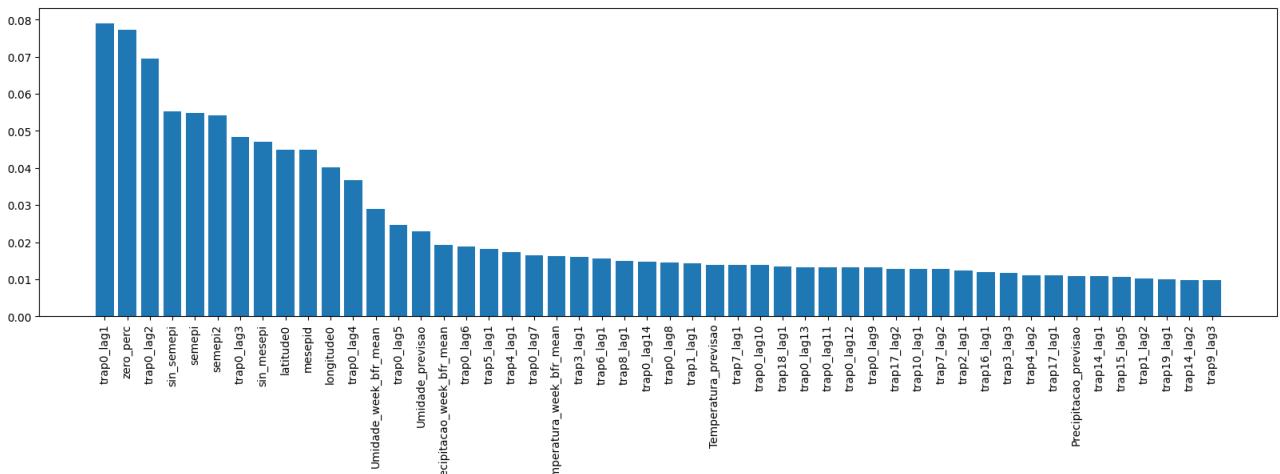
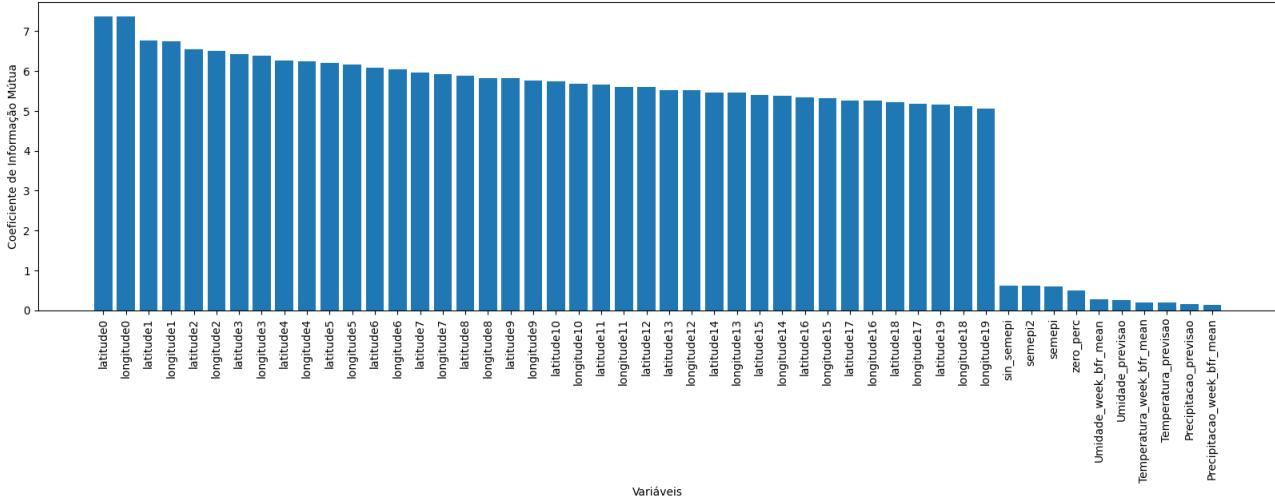
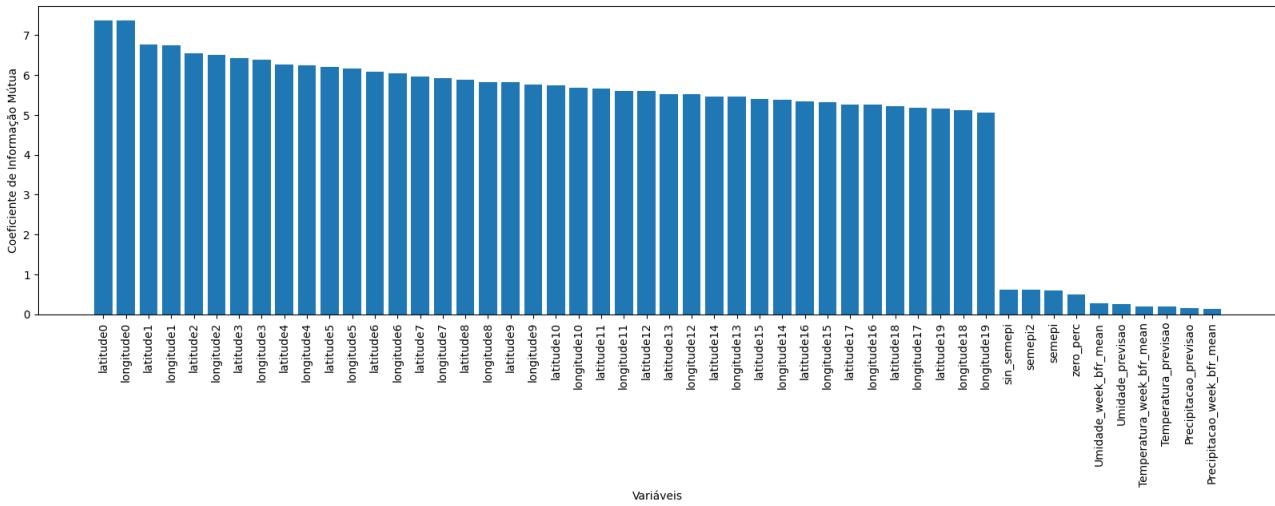


Figura 25: Informação mútua entre entradas e número de ovos - 14 lags, 20 vizinhos

demais coordenadas é consideravelmente alta, como pode ser visto na Figura 26. A correlação entre o primeiro par de coordenadas também é alta, entretanto, descartar uma das componentes comprometeria a informação espacial da amostra. Assim, o par foi inteiramente conservado.



(a) Longitude



(b) Latitude

Figura 26: Informação Mútua entre as coordenadas da armadilha principal e das demais armadilhas

A análise anterior foi repetida para o conjunto de dados com o $l = 5$ e $t = 10$, resultando nos novos coeficientes da Figura 27. Valores menores que 0.1 foram filtrados para melhorar a visualização. Comportamento semelhante ao encontrado previamente foi observado, ressaltando ainda mais a baixa correlação entre a saída e as coletas das armadilhas vizinhas com atrasos maiores que 1. Novamente, para manter algum grau de informação espacial, decidiu-se por utilizar as 9 armadilhas vizinhas, entretanto, com apenas dois valores passados. Com essas escolhas, o conjunto de dados final continha 282169 amostras, das quais 60.2 % eram nulas. No conjunto de teste derivado dele, as placas vazias compreendiam 55.2%. Em relação ao problema de três classes, proporções entre as classes semelhantes às iniciais foram encontradas.

Neste momento, avaliou-se também a importância dos dados meteorológicos, e se haveria diferença em utilizar os períodos de tempo distintos. Como pode ser aferido, as variáveis exógenas apresentam correlação relevante em relação ao número de ovos, não havendo prevalência clara de um período sobre o outro. A semana anterior à instalação foi escolhida para os próximos testes por conta de sua disponibilidade, evitando o uso de variáveis meteorológicas preditas por entidades externas.

A Figura 28, por sua vez, apresenta os resultados da análise anterior, porém para o vetor booleano indicando a presença de ovos na armadilha. Apesar da mudança na ordem das variáveis mais correlacionadas, o grupo é o mesmo relatado para a análise com a contagem de ovos.

A última relação avaliada por este método foi entre as variáveis temporais e suas transformações.

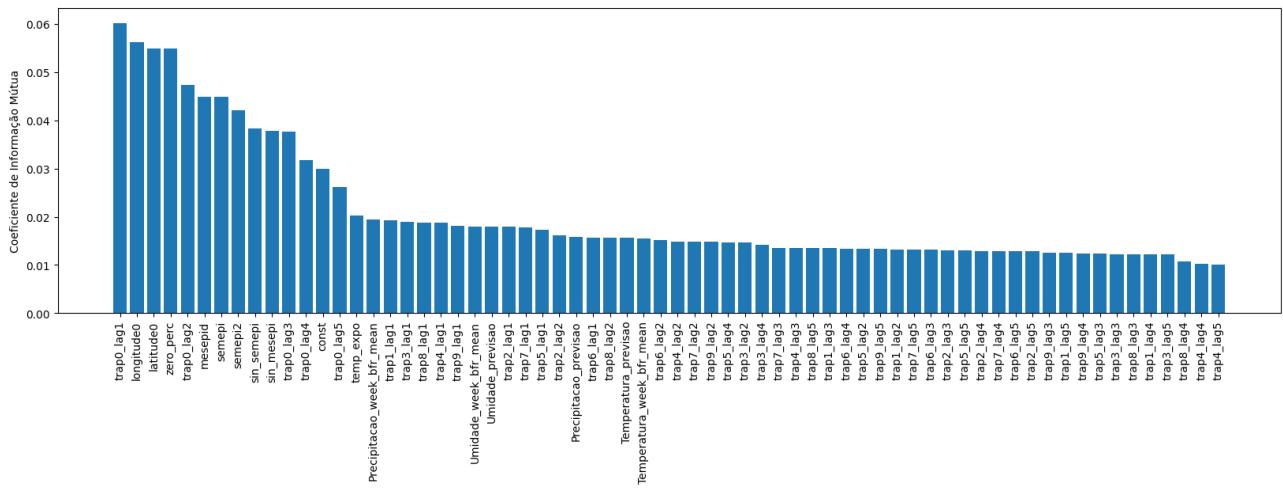


Figura 27: Informação mútua entre entradas e número de ovos - 5 lags, 10 armadilhas

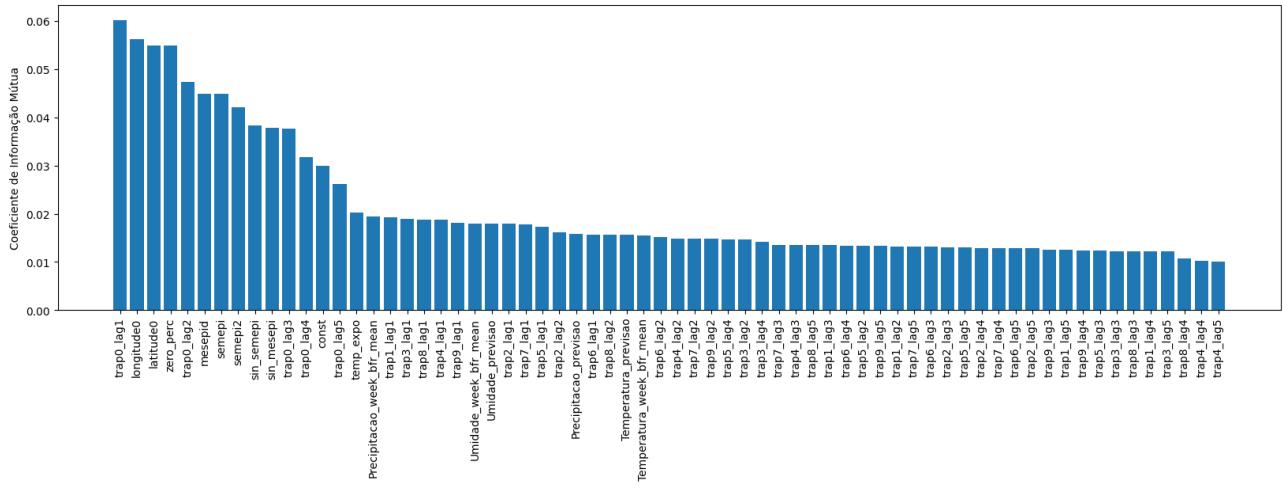


Figura 28: Informação mútua entre entradas e série indicando presença de ovos - 5 lags, 10 armadilhas

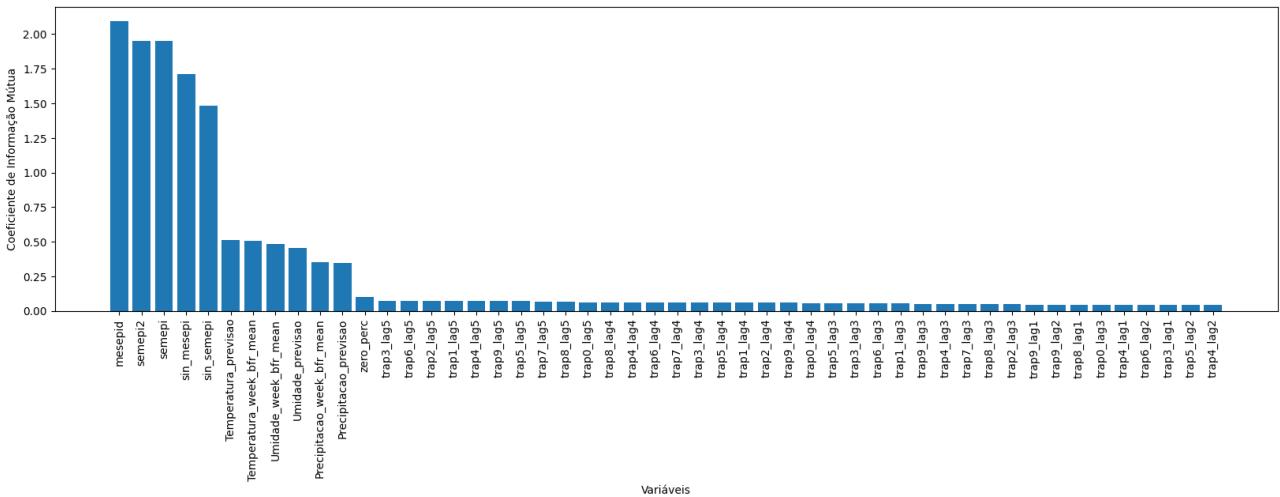


Figura 29: Informação mútua entre entradas e mês epidemiológico - 5 lags, 10 armadilhas

O mês epidemiológico foi escolhido como referência por ser a variável mais simples e preferida em casos de redundância. Os coeficientes apresentados na Figura 29 indicam de fato alta sobreposição de informação.

Em corroboração com a aplicação da Informação Mútua, o método baseado no Critério de Informação de Akaike (AIC) foi aplicado ao problema de classificação de presença para os dados com $l = 5$ e $t = 10$. As três políticas resultaram em modelos com acurárias no conjunto de teste semelhantes (70.0%), embora o número de variáveis tenha sido alterado consoante as políticas. O modelo final da política 'backward' apresentou dez variáveis a menos que os demais, sendo elas todas as cinco amostras atrasadas da armadilha principal, as duas amostras anteriores às vizinhas, as coordenadas, a porcentagem de zero e o mês epidemiológico. Os modelos resultantes dos outros processos tinham por adição amostras mais antigas das armadilhas vizinhas, as coordenadas de algumas dessas armadilhas e demais variáveis temporais.

Resultado semelhante foi observado ao desenvolvemos os modelos de referência com as diferentes estruturas temporais e espaciais propostas. Todos apresentaram desempenhos similares, com acurárias próximas de 70.0% e RMSE de 32.7 ovos. Consequentemente, optou-se por utilizar o mês epidemiológico e as coordenadas da armadilha principal como representações das respectivas estruturas.

Nesse contexto, os modelos referências foram treinados novamente, utilizando agora as variáveis escolhidas com exceção das variáveis exógenas, no intuito de avaliar a sensibilidade dos resultados a sua presença. A redução na acurácia, em comparação com o modelo logístico completo, foi de apenas 0.3%, enquanto o RMSE aumentou 0.1 ovos. Dada essa invariância, optou-se por, em definitivo, desconsiderar as informações meteorológicas dos dias de exposição da placa de interesse e utilizar os dados referentes à semana anterior.

Enfim, tendo sido escolhidos o grupo final de variáveis de entrada, o processamento dessas foi definido por meio do mesmo procedimento com os modelos de referência. Dentre todos os modelos, os com melhor acurácia foram os que truncavam as contagens de armadilhas de entrada em 100. As melhorias foram de 0.3% e 0.1 ovos em relação aos modelo sem processamento. Portanto, a entrada escolhida para os próximos experimentos tiveram as variáveis entomológicas saturadas em 100. Os resultados para todos os modelos descritos podem ser encontrados no Apêndice ??

Em resumo, após a seleção das entradas e a análise do impacto de cada uma delas, bem como de seu processamento nos resultados dos modelos de referência, foi definido que a entrada para os modelos finais incluiria as seguintes variáveis:

- Os cinco valores da armadilha analisada imediatamente anteriores à amostra principal, saturados em 100

- Os dois valores imediatamente anteriores à amostra principal das nove armadilhas disponíveis mais próximas à principal, saturados em 100
- As coordenadas da armadilha principal
- A média das variáveis meteorológicas na semana anterior à instalação da placa analisada
- A porcentagem de zeros na armadilha principal até o momento
- O mês epidemiológico transformado em variável *dummy*

Todas as entradas foram normalizadas.

5.2 Resultados dos Modelos Preditivos

Definidas as entradas e o processamento, os modelos finais foram treinados.

Sumarizado na tabela 5 estão os resultados dos melhores exemplares dos modelos testados para os três problemas propostos. Entre colchetes encontram-se os resultados equivalentes para os dados de treinamento.

Modelo	Régressão (RMSE)	Classificação da Presença (Acurácia)	Classificação nas Faixas (Acurácia)
Naive	38.7 [35.4]	66.3% [67.7%]	65.2% [68.1%]
Linear	32.7 [27.3]	70.3% [72.7%]	72.5% [75.5%]
Multi-Layer Perceptron	48.0 [38.9]	70.4% [72.7%]	72.6% [75.8%]
Support Vector Machine	42.3 [36.9]	69.8% [72.4%]	71.6% [74.7%]
Random Forest	30.2 [25.4]	70.8% [77.2%]	72.5% [77.1%]
CatBoost	29.9 [26.8]	70.8% [73.3%]	72.7% [75.9%]

Tabela 5: Desempenho dos melhores modelos de cada família para os três problemas, aplicados no grupo de teste e, em chaves, de treinamento.

Embora os modelos Catboost tenham apresentado os melhores resultados para todos os problemas analisados, não houve alteração significativa nos resultado com o aumento da complexidade dos modelos em comparação aos modelos de referência. Curiosamente, nenhum dos modelos de classificação das faixas previu valores para a classe 'média', a menos representativa neste problema desbalanceado. Como tentativa de contornar esse comportamento, foram atribuídos pesos maiores às amostras dessa classe durante o treinamento do modelo logístico. Apesar de, com essa abordagem, os modelos apresentarem previsões para a classe em questão, a acurácia obtida foi semelhante ao modelo ingênuo.

Os resultados detalhados dos modelos testados, assim como seus hiperparâmetros, são descritos no Apêndice ??.

6 Discussão

Dadas as proporções do problema, os resultados obtidos são, no melhor dos casos, medíocres. A melhoria obtida em relação aos modelos ingênuos pode ser considerada irrisória e o erro, em média de 30 ovos em um universo de 100 e de 30% das classificações, torna inviável a aplicação dos modelos em situações práticas. Nesse sentido, as razões para esses resultados foram investigadas. A complexidade e diversidade de modelos testados sugerem, a princípio, ausência de capacidade preditiva nos dados de ovitrampas, na medida em que a distribuição de ovos não estaria relacionada com as variáveis empregadas a ponto de podermos prever novas contagens com precisão aceitável,

pelo menos considerando a resolução espaço-temporal adotada. Outra hipótese para esse desempenho seria a excepcionalidade dos anos testados, conforme relatado pelo corpo técnico da prefeitura, dado que, os anos de 2022_23 e 2023_24 apresentaram mudanças no perfil das posturas em relação aos anos anteriores.

Sob essa ótica, os resíduos dos modelos com as melhores métricas, os Catboosts, foram analisados em detalhes na tentativa de identificar padrões. Além disso, novos modelos foram treinados com o objetivo de explorar as diferenças entre os dados anuais. O foco dessa subseção recai sobre os problemas de classificação de presença e previsão da contagem em detrimento do problema de classificação das faixas de valor.

6.1 Catboost para Classificação de Presença

Embora a acurácia do modelo Catboost para o problema de classificação de presença tenha sido de 70.8%, sua precisão é distinta para cada classe, como verificado na Matriz de Confusão 6. Nota-se a tendência do modelo de prever a classe ausente em detrimento da outra, de modo que sua sensibilidade é consideravelmente menor que sua especificidade, respectivamente, 58.6% e 80.8%. Foram previstos como 'Ausente' 63.2% das amostras, valor muito acima dos 55.2% de amostras vazias no conjunto de testes. O percentual de ausências nos dados de treinamento, entretanto, foi de 60.2%. Isso indica que mudanças nas dinâmicas dos dados ao longo dos anos poderiam ser causas para o desempenho dos modelos, configurando um exemplo de Concept Drift (49)

	Anterior: Ausente	Anterior: Presente
Atual: Ausente	44.6%	10.6%
Atual: Presente	18.6%	26.3%

Tabela 6: Matriz de Confusão Percentual do Catboost para classificação da presença de ovos

Baseado nesta hipótese, os anos de treinamento e de teste do modelo logístico foram alterados. Todos os anos epidemiológicos disponíveis foram utilizados tanto como treinamento e teste em pares, independente da ordem cronológica. A matriz 30 condensa os resultado desses modelos. Em seu eixo x encontram-se os anos utilizados para treinamento do modelo logístico, enquanto no eixo y estão os anos de teste. Sua diagonal foi preenchida com os resultado do modelo ingênuo daquele ano. A variância dos resultados encontrados foi de aproximadamente 10%. Nota-se, na matriz, padrões acentuados em suas linhas horizontais em comparação às verticais, indicando a preponderância dos anos de teste para o resultado do modelo em relação aos anos de treinamento. Dados os mesmos anos de teste, os modelos tiveram consistentemente acurácia 5% maior que a do modelo ingênuo. Os últimos anos epidemiológicos da base de dados mostraram-se difíceis de serem previstos em relação ao restante dela, sendo 2023_24 o com piores resultados dentre os anos completos. Outro experimento realizado, uma espécie de validação cruzada, considerou todos os anos epidemiológicos como treino, com exceção de um, que seria o de teste. O resultado, exposto no gráfico 31, apresenta perfil semelhante à da matriz. Em adição à acurácia, foi disposta no gráfico a porcentagem de zeros no ano de teste. A correlação de Pearson entre suas séries de diferenças foi de 0.861. O vazamento de dados, existente nas entradas relativas à porcentagem histórica de zeros na armadilha e aos valores atrasados das amostras iniciais, aparentemente pouco influenciou os resultados dada a simetria horizontal da matriz 30 em relação à diagonal.

Apesar das diferenças encontradas com a alteração dos dados de treinamento e teste, nenhum dos modelos obteve acurácia superior a 80%. Há, portanto, um grupo de amostras consistente que nunca é acertado. Por isso, na tentativa de encontrar algum padrão nos dados incorretos, os resíduos do modelo final foram marginalizadas conforme o ano epidemiológico, o mês epidemiológico e as categorias da prefeitura (Tabela 7).

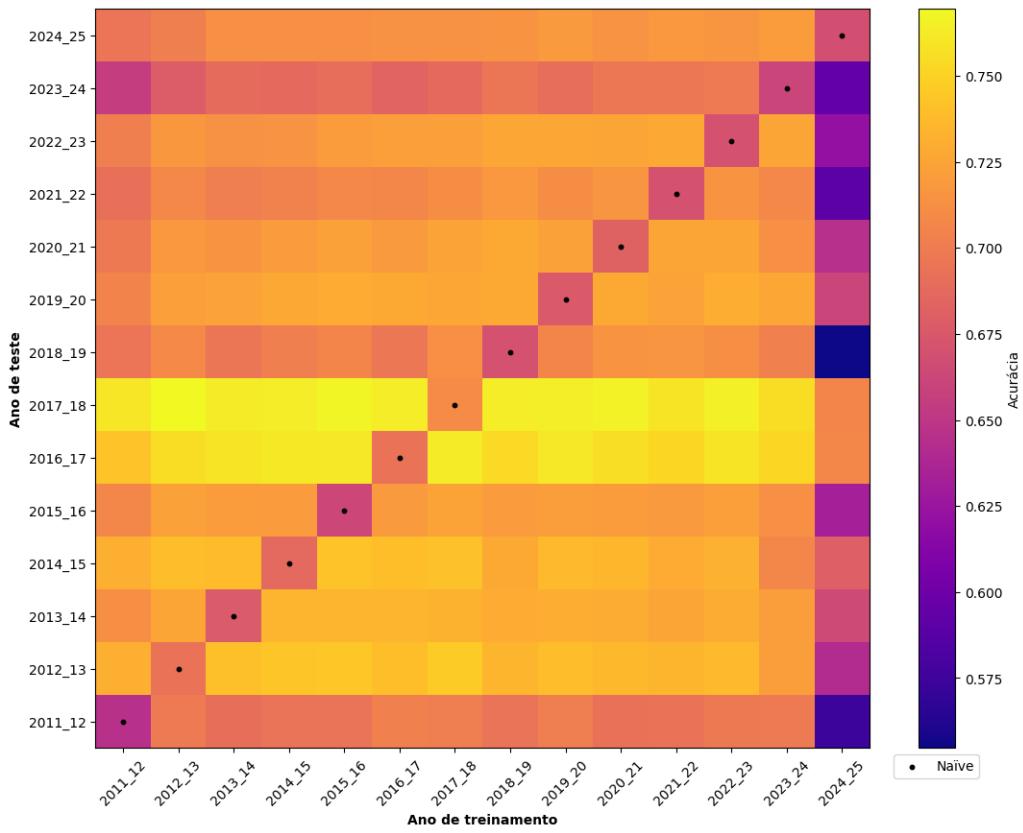


Figura 30: Acurácia de modelos logísticos treinados e testados com diferentes anos epidemiológicos. Na diagonal, foram posicionados o resultado dos modelos Naïve

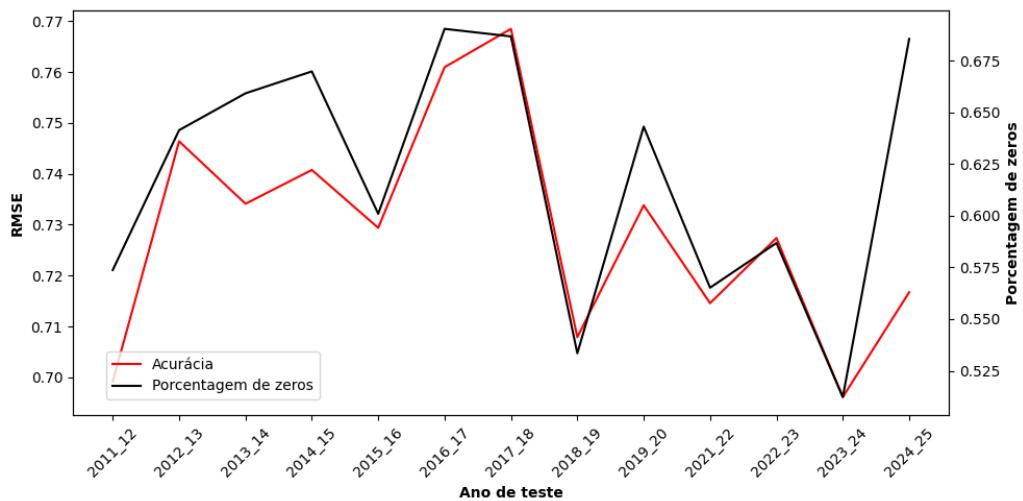


Figura 31: Em vermelho, a acurácia do modelo logístico utilizando os anos epidemiológicos do eixo x como teste e os demais como treinamento. Em preto, a porcentagem de zeros no ano de teste

Os erros em cada ano epidemiológico seguiram perfil semelhante ao encontrado na análise de validação cruzada da Figura 31, com a porcentagem de erros por ano comportando-se de forma praticamente uniforme, apesar da diferença na porcentagem de zeros desses anos. Os erros condicionados ao mês epidemiológico variaram entre si em até 13 pontos percentuais, sem, entretanto, apresentar padrão referente à porcentagem de zeros verificado no Gráfico 16b. Comportamento semelhante ocorreu entre as Categorias, com grande variância, porém, sem que o erro seguisse a porcentagem de zeros. A categoria B, com a maior porcentagem de valores nulos, apresenta o menor percentual de erros. No entanto, ao contrário do indicado nas análises anteriores, a segunda categoria com o menor percentual de erro é a A1, a com o menor percentual de amostras vazias. Isso sugere que o desbalanceamento entre as classes, e não o percentual de zeros, é o principal fator responsável pela melhoria nos resultados do modelo de classificação.

Para um estudo espacial dos erros, agrupou-se no Mapa 32 os resíduos conforme sua armadilha. Nele, o raio das armadilhas é proporcional à porcentagem de erros e suas cores indicam a faixa percentual a qual ela pertence, segundo essa métrica. A distribuição espacial dos erros é aparentemente semelhante ao número médio das contagens, disposto no mapa 11, tendo a zona sudeste da cidade concentração de armadilhas com menores percentuais de erro e as zonas central e norte concentração de armadilhas com maior percentuais.

Ano Epidemiológico	Mês Epidemiológico	Categoria
2022_23 2023_24 2024_25	Junho	29.3%
	Julho	29.2%
	Agosto	25.7%
	Setembro	29.4%
	Outubro	34.2%
	Novembro	27.8%
	Dezembro	21.1%
	Abri	31.8%
	Maio	32.6%
A1 A2 M B		26.7%
		30.9%
		28.5%
		22.4%

Tabela 7: Porcentagem de condicionada a Ano Epidemiológico, Mês Epidemiológico e Categoria do modelo de Classificação

6.2 Catboost para Regressão

Análises análogas às descritas na subseção anterior foram realizadas para o modelo Catboost aplicado ao problema de regressão.

A matriz 33, que mostra o RMSE dos modelos com os diferentes pares de anos de treinamento e teste, apresenta resultado semelhante à equivalente para o problema de classificação, com padrões acentuados nas linhas horizontais e distâncias consistentes entre os resultados dos modelos ingênuos e os demais modelos de cada ano de teste. Pontua-se que a coloração da escala foi invertida, em conformidade com a interpretação das métricas utilizadas.

Já o gráfico 34, com as séries dos RMSE da análise de Validação Cruzada, segue o perfil semelhante à matriz 33 e ao gráfico da subseção anterior. A correlação entre o RMSE e a porcentagem de zeros continua alta, porém é negativa (-0.928), resultado que reforça a importância da porcentagem de zeros para os modelos preditivos.

Em relação aos agrupamentos por ano e mês epidemiológicos e categoria (Tabela 8), os resultados apresentaram diferenças em relação aos seus similares. Enquanto se nota uniformidade entre as acuráncias anuais na análise da classificação, os erros médios dos modelos regressivos foram mais dispare. Os valores mensais também apresentam padrões distintos, com a mudança dos meses com

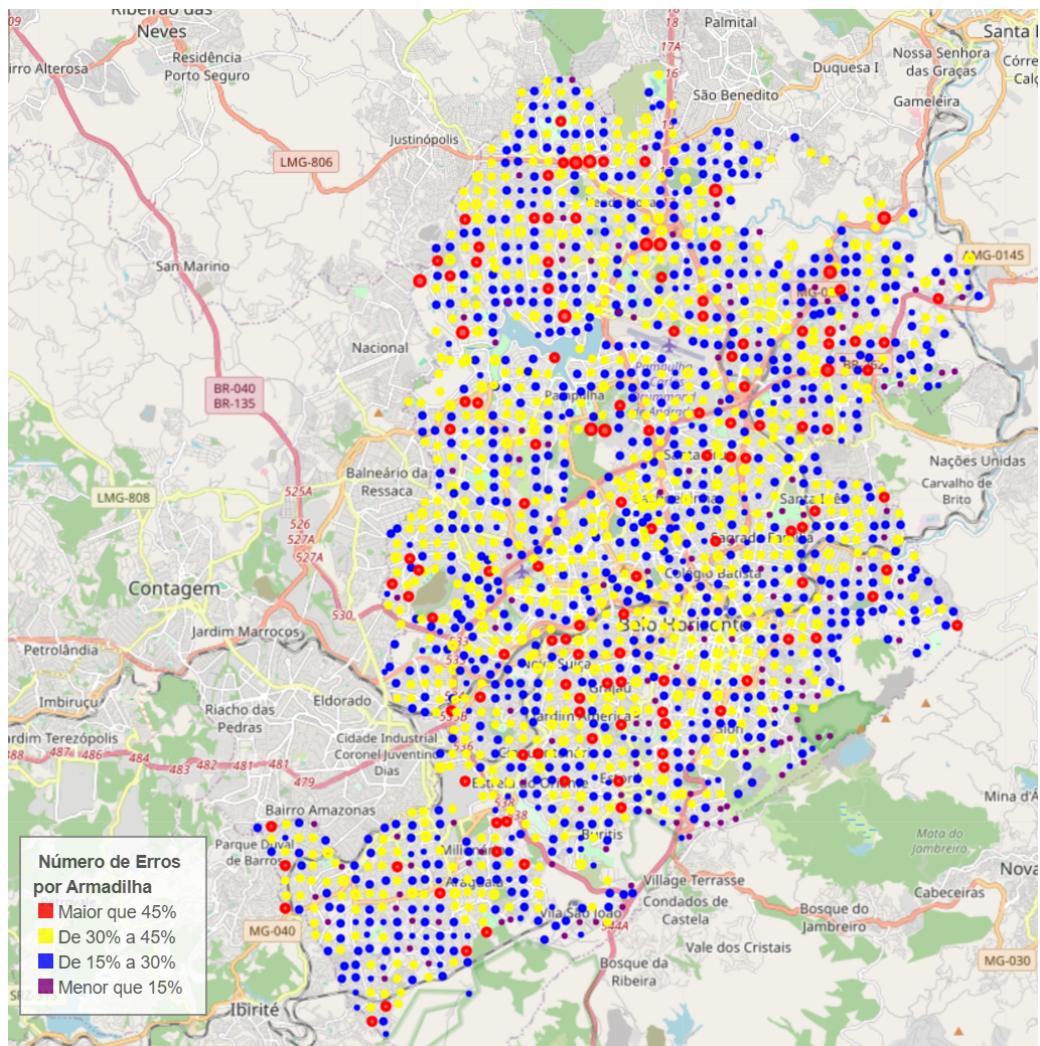


Figura 32: Porcentagem dos erros por armadilha do melhor modelo de Classificação. As cores dos ícones indicam faixa à qual a porcentagem de erros pertence e os raios são proporcionais ao valor em si

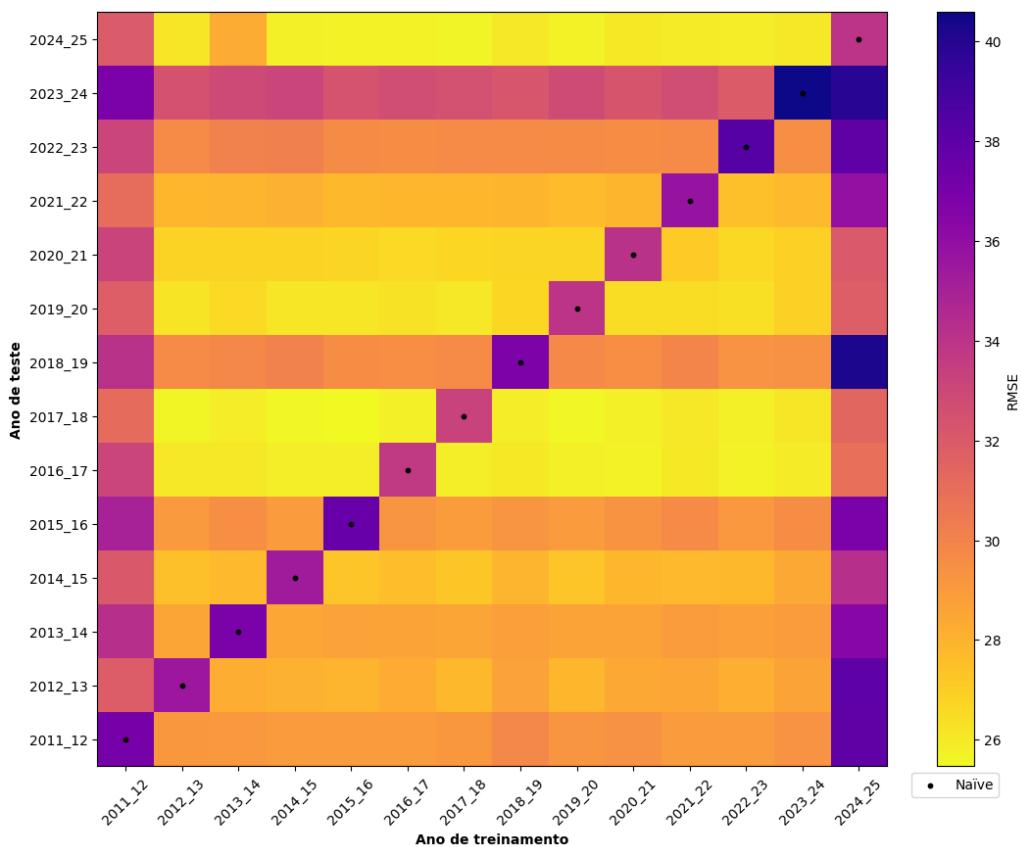


Figura 33: RMSE de modelos lineares treinados e testados com diferentes anos epidemiológicos. Na diagonal, foram posicionados o resultado dos modelos Naïve

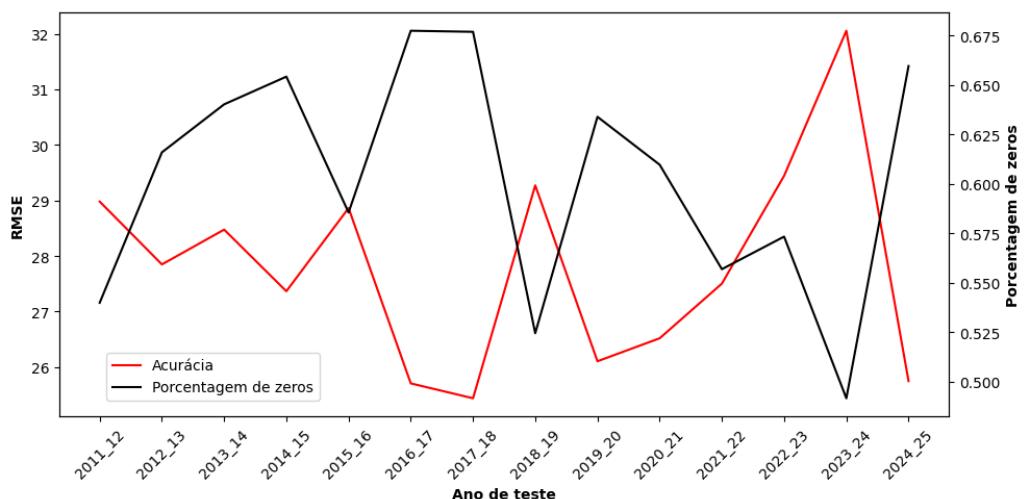


Figura 34: Em vermelho, o RMSE do modelo linear utilizando os anos epidemiológicos do eixo x como teste e os demais como treinamento. Em preto, a porcentagem de zeros no ano de teste

maior erro relativo. Entretanto, ao contrário da análise anterior, nota-se nesta clara relação entre os meses com maior número de ovos média e os meses com maior erro. Diferença semelhante ocorreu na análise das categorias, em que a classe A1 deixou de ser a segunda classe menos errada para se tornar a com maior erro médio.

O mapa 35, do RMSE por amostra, apresenta regiões mais bem delimitadas, em relação ao mapa 32, aproximando-se ainda mais do padrão visto no mapa 11. Segue o mesmo padrão de cores e raios para a esquematização das armadilhas.

Ano Epidemiológico	Mês Epidemiológico	Categoria
2022_23 20.543952 2023_24 24.895627 2024_25 17.968624	Junho	18.828738
	Julho	16.965916
	Agosto	16.083753
	Setembro	19.731593
	Outubro	27.019493
	Novembro	30.031129
	Dezembro	30.348795
	Abril	29.779346
	Maio	24.254281

Tabela 8: Porcentagem de Resíduos agrupadas por Mês Epidemiológico, Ano Epidemiológico e Categoria do modelo de Classificação

Em adição às análises anteriores, quatro amostras de cada uma das categorias foram escolhidas randomicamente e a série temporal de seus valores foi disposta junto aos valores previstos pelo modelo, Figuras 36, 37, 38 e 39

Salvo algumas amostras excepcionais, para todas as classes, os modelos comportam tipicamente como um modelo com forte influência da variável autorregressiva, em que a tendência de mudança de seu resultado é definida pelo valor anterior. Entretanto, nota-se dificuldade de prever aumentos ou quedas bruscas das contagens, o que explica o motivo das armadilhas da categoria A1, com variações muito acentuadas, apresentarem o pior percentual de erros. Outro comportamento aferido é a dificuldade do preditor em acompanhar as séries de zeros. Ele, entretanto, é característico das armadilhas de baixa média de ovos (classe B), e, por conseguinte, não pode ser associado aos altos valores dos resíduos.

Em síntese, não foi encontrando nenhum padrão que explicasse os erros além do balanceamento das classes para o modelo classificador e da variância dos valores para o modelo de regressão. Tampouco foi encontrada diferença nos dados a ponto de explicar os desempenhos dos modelos experimentados. Com base nessas análises, conclui-se que os dados de fato não apresentam poder preditivo para a resolução espaço-temporal analisada, em corroboração com a literatura disponível.

7 Conclusão

Este Trabalho de Conclusão de Curso teve por objetivo estudar a rica base de dados de ovitrampas coletada pela Prefeitura de Belo Horizonte (PBH) e avaliar sua capacidade preditiva, considerando alta resolução espaço-temporal. Aliados à introdução de variáveis meteorológicas coletadas na região metropolitana da cidade e a métodos simples de processamento de dados, modelos foram treinados na tentativa de prever três características de amostras futuras: a presença de ovos, a contagem de ovos e a faixa na qual essa contagem se posicionaria, uma simplificação do problema anterior. Apesar do aumento incremental na complexidade dos modelos testados, pouca melhoria foi observada nas métricas avaliadas. Os modelos finais foram levemente superiores ao modelo ingênuo, aquele que apenas repete a amostra anterior, com resultados insuficientes para serem aplicados em cenários reais.

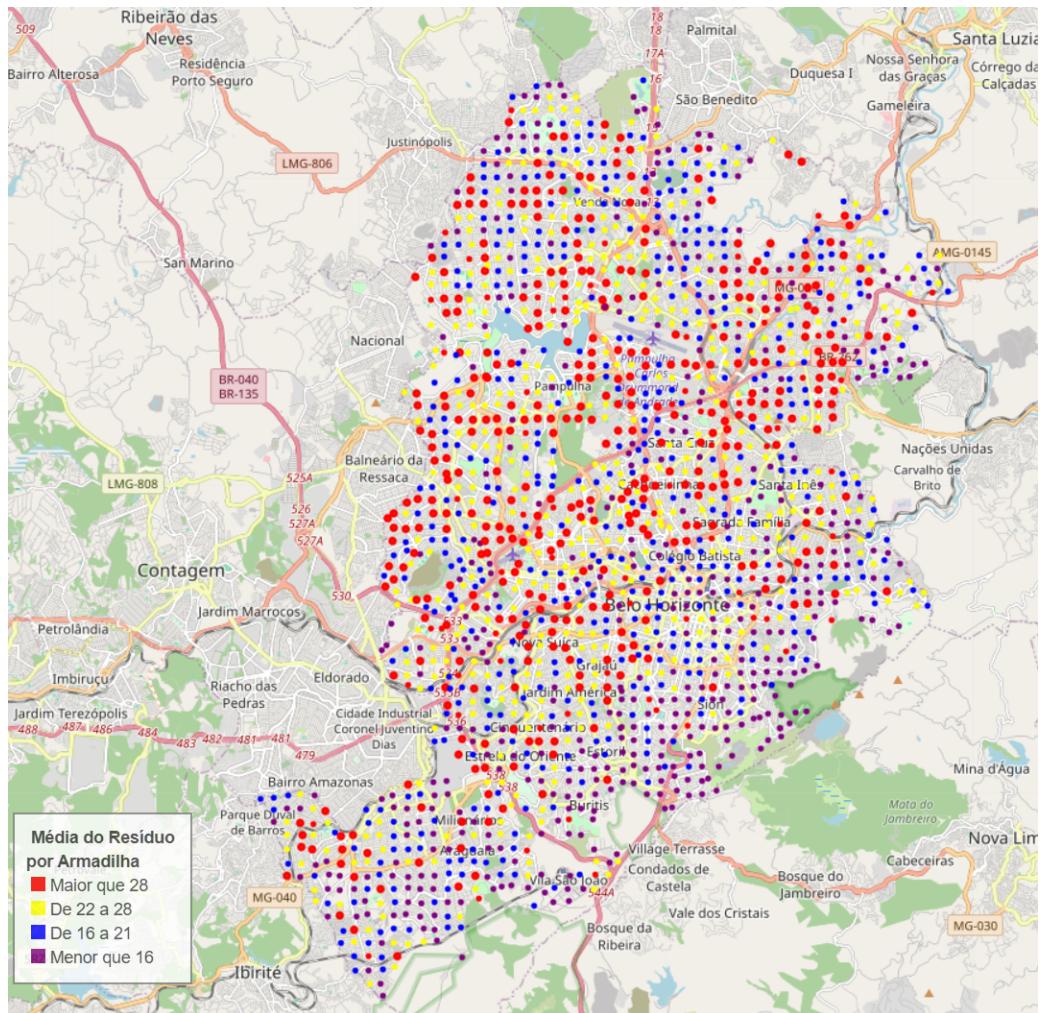


Figura 35: Média dos resíduos do melhor modelo de Regressão por armadilha. As cores dos ícones indicam o quartil a qual pertencem e os raios são proporcionais aos valores

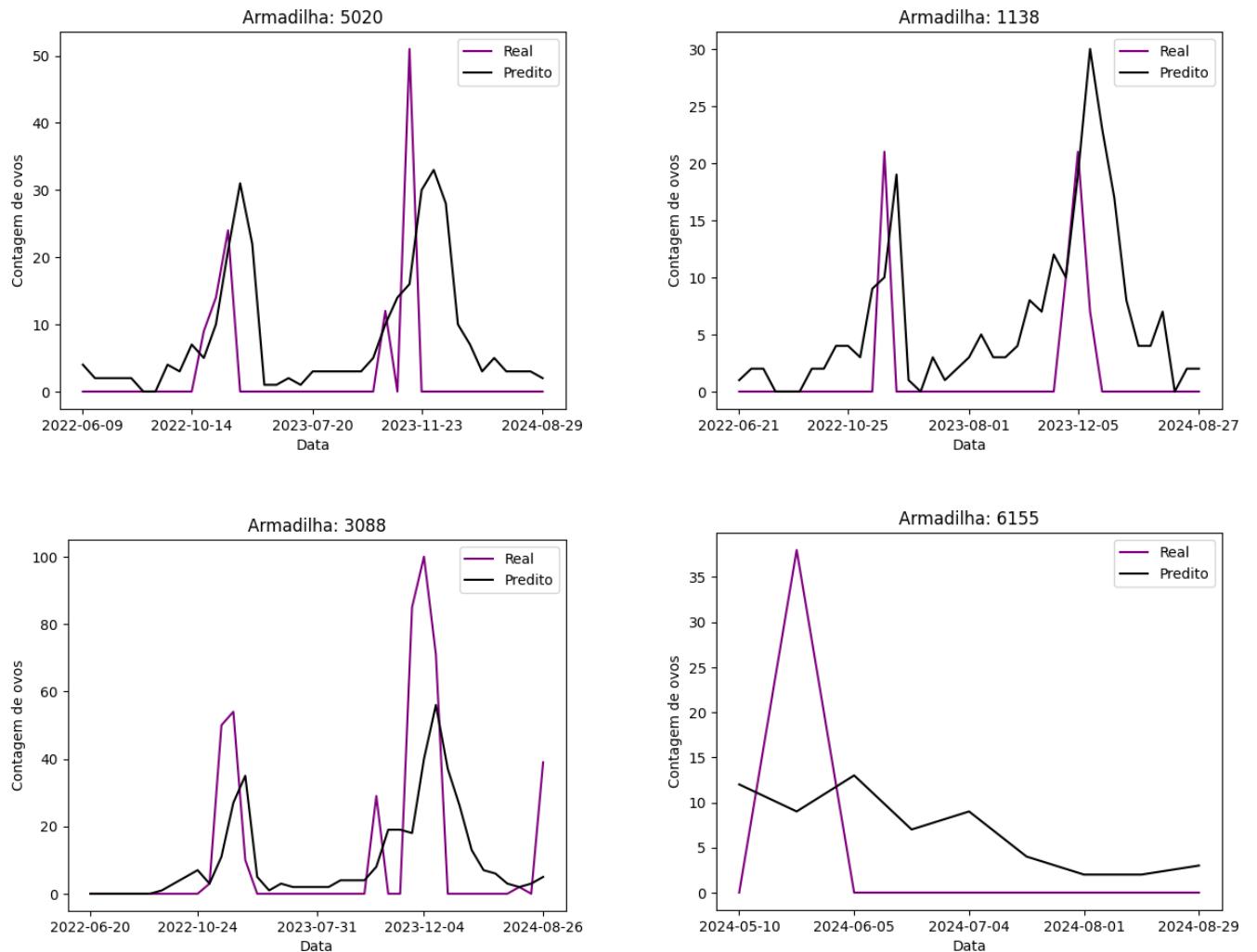


Figura 36: Séries dos valores reais e preditos de quatro armadilhas da Categoria B

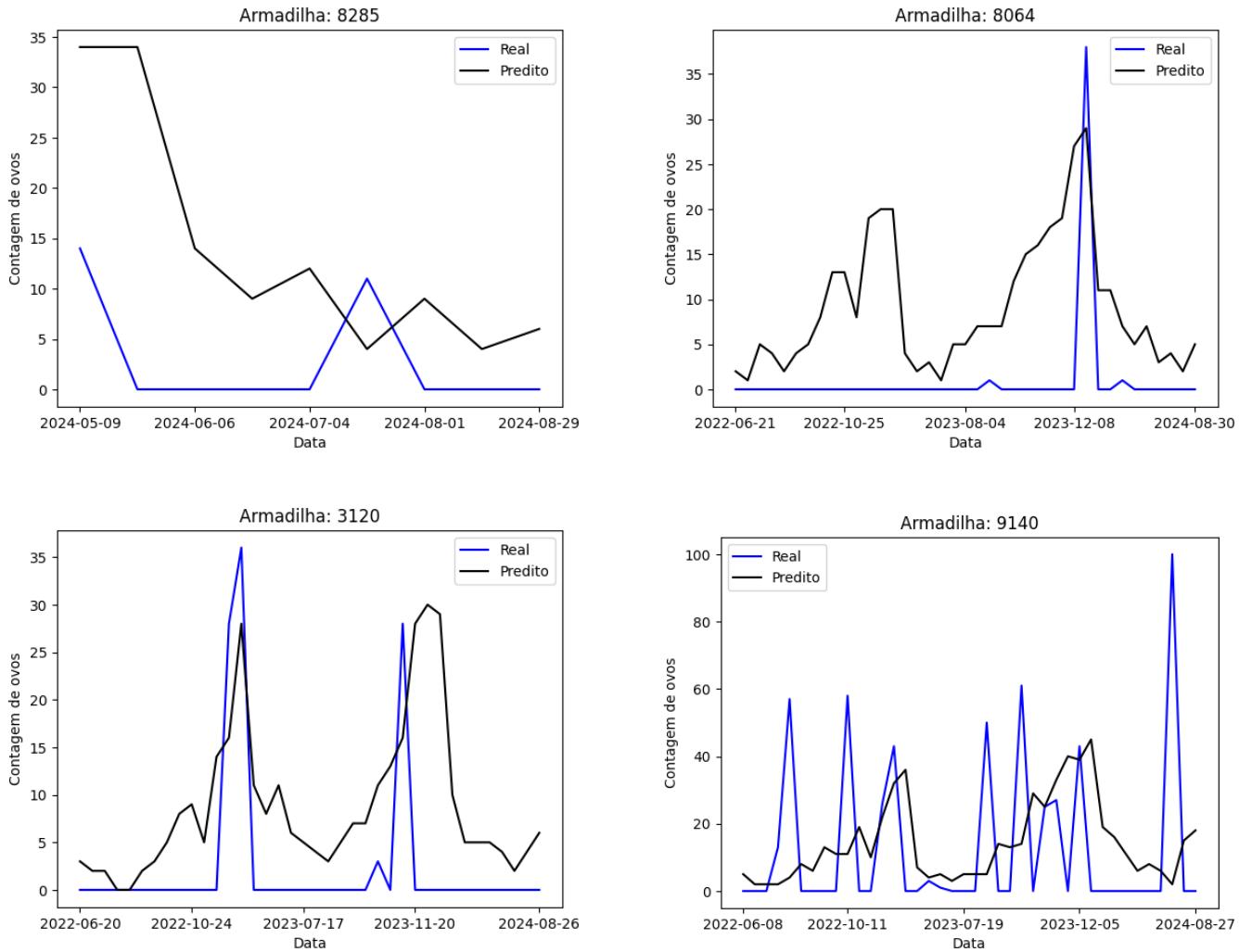


Figura 37: Séries dos valores reais e preditos de quatro armadilhas da Categoria M

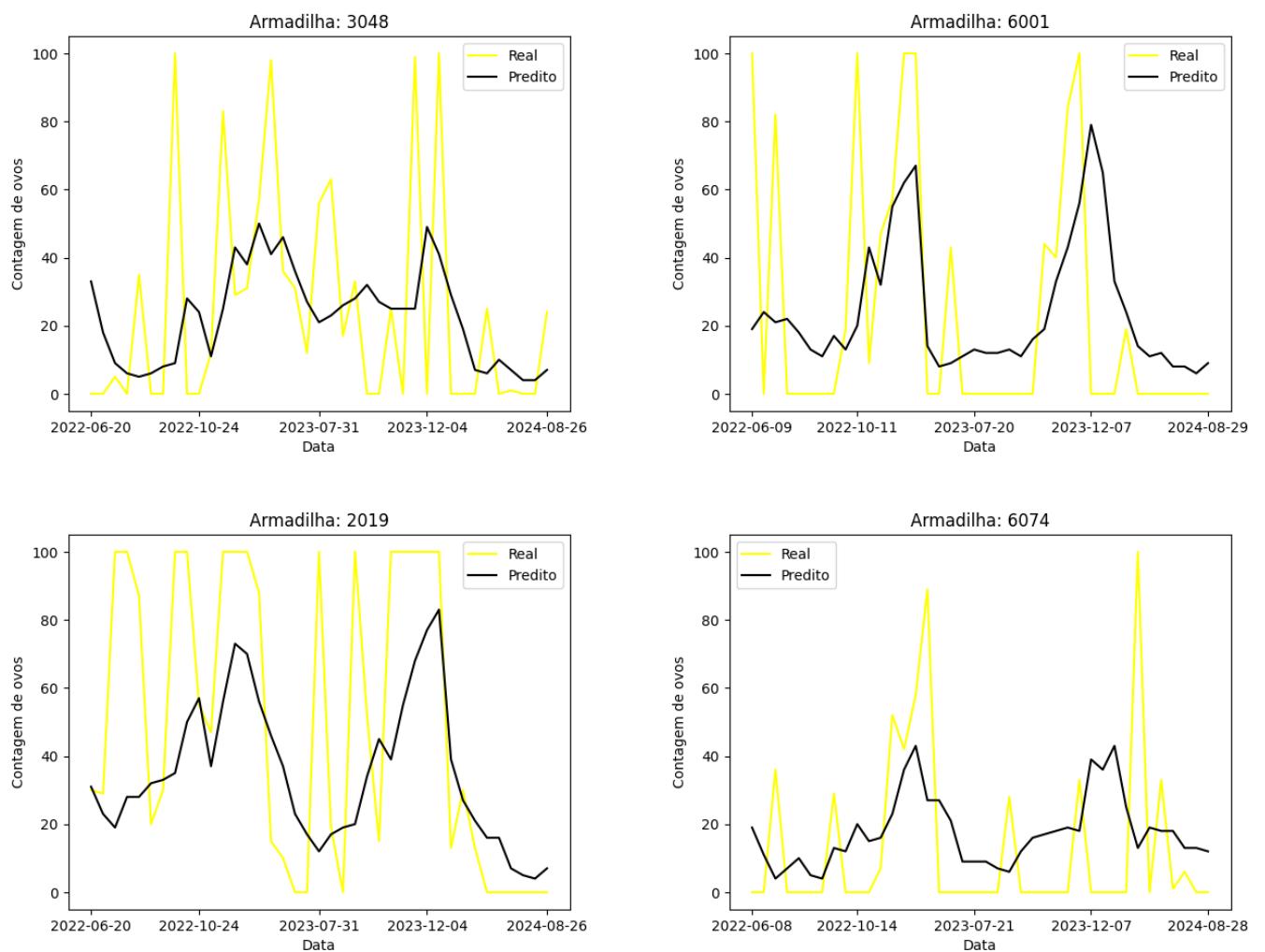


Figura 38: Séries dos valores reais e preditos de quatro armadilhas da Categoria A2

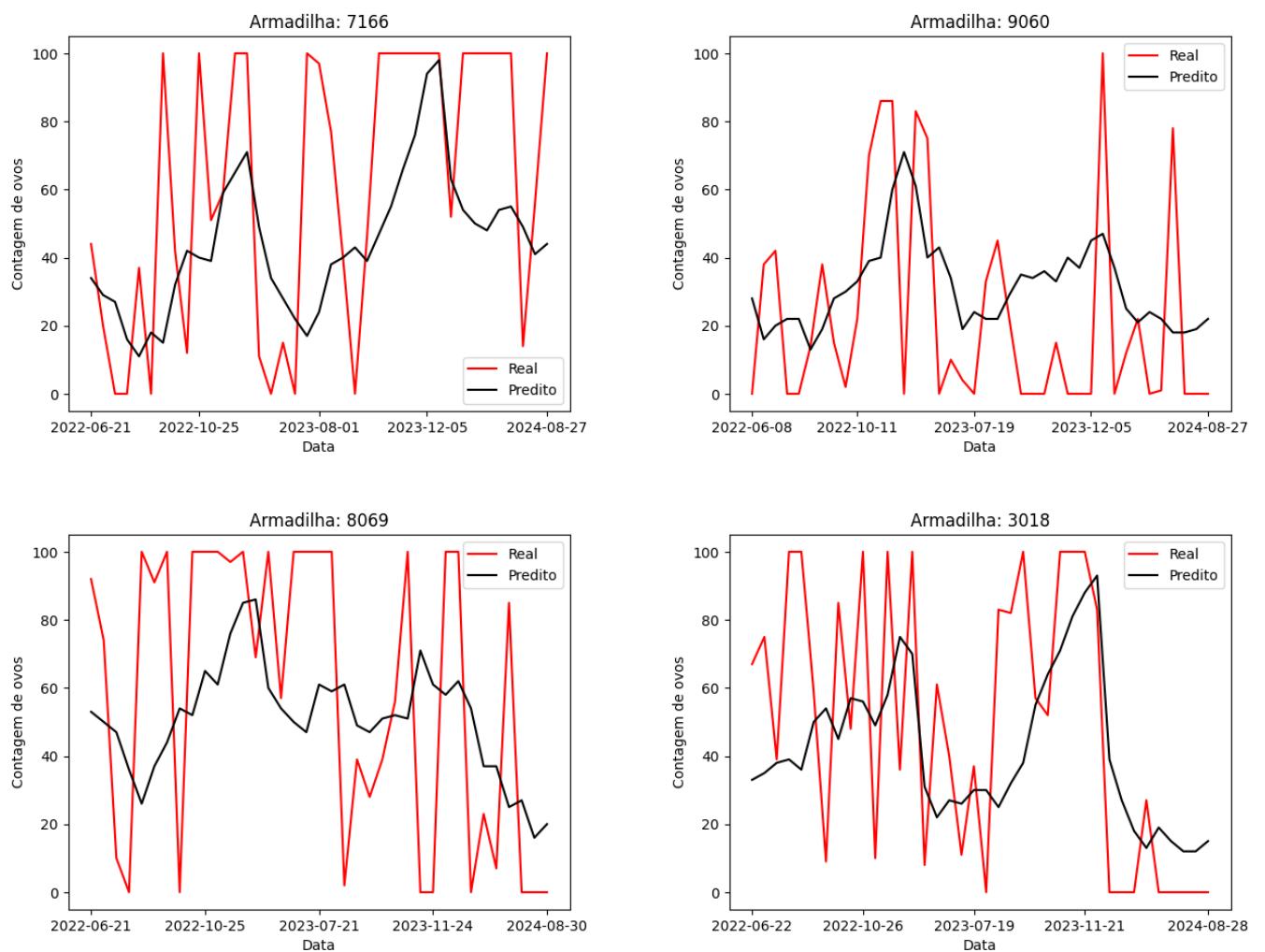


Figura 39: Séries dos valores reais e preditos de quatro armadilhas da Categoria A1

Os resíduos do melhor modelo treinado para os problemas de regressão e classificação, o Catboost, foram analisados na busca por padrões que explicassem o desempenho relatado. Além da própria variação dos dados, não foram encontrados padrões espaciais ou temporais consideráveis.

Apesar de tais resultado indicarem para uma ausência da capacidade preditiva nesses dados, comprehende-se que novas análises e melhorias nos métodos podem ser feitas. Estruturar melhor os dados espaciais e temporais das armadilhas seria um primeiro passo. Na dimensão temporal, por exemplo, isso se daria pela adição de novas variáveis aos modelos com informações de longo prazo da armadilha, como a média histórica de sua coleta ou a própria categoria gerada pela PBH, que não foi adotada por ser considerada uma etapa de processamento custosa. Já na dimensão espacial, a escolha dos vizinhos adicionados pelo modelo pode-se dar por meio de critérios diferentes da distância, como a correlação entre as séries históricas ou por estrutura mais complexas de agrupamento, como o Estimador de Densidade de Kernel (KDE) (7) ou pelo algoritmo SKATER (2). Reconhece-se que as informações das caudas foram pouco exploradas na criação dos modelos finais. Isso ocorreu porque as tentativas iniciais resultaram em modelos com métricas não apenas piores que os modelos ingênuos, mas também piores do que uma escolha completamente aleatória. Ainda assim, as amostras com valores extremos são de grande relevância para a compreensão da dinâmica dos mosquitos, e esforços para incorporá-las podem gerar bons resultados. O processamento dos dados meteorológicos pode igualmente ser aprimorado ao considerar também seus valores máximos e mínimos, utilizar dados diários em vez de agregados semanalmente, ou ainda considerar valores globais relativos a toda a região municipal. Ademais, podem ser incorporadas ao processamento dessas variáveis o conhecimento do ciclo biológico do mosquito, priorizando os valores nos horários de maior atividade dos vetores ou integrando a interferência da temperatura e pluviosidade nos tempos entre postura de ovos por um mesmo mosquito, conforme indicado pela literatura (8). Outras variáveis exógenas comuns na literatura, como a altitude (8) ou a densidade(4) populacional do local da armadilha, podem ser adicionadas à modelagem, adicionando informações inexistentes.

Considerando a riqueza da base de dados trabalhada, entendemos que a análise preditiva realizada é apenas uma das facetas a serem exploradas. A identificação prévia de anos de epidemia e a comparação das dinâmicas de postura nesse período, por exemplo, são outras abordagens possíveis, que agregam valor para a condução de políticas públicas no combate às arboviroses. Neste sentido, ressalta-se que este projeto é apenas o início de uma colaboração entre a Prefeitura de Belo Horizonte e a Universidade Federal de Minas Gerais.

8 Bibliografia

Referências

- [1] Luis Antonio Aguirre. *Introduçaoa identificaçao de sistemas–Técnicas lineares enao-lineares aplicadas a sistemas reais*. 2015.
- [2] Renato M Assunção, Marcos Corrêa Neves, Gilberto Câmara, and Corina da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.
- [3] Renato M. Assunção. *Statistical Foundations for Data Scientist*. 2022.
- [4] Aswi Aswi, SM Cramb, Paula Moraga, and Kerrie Mengersen. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology & Infection*, 147:e33, 2019.
- [5] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

- [6] Anthony J Bishara and James B Hittner. Testing the significance of a correlation with non-normal data: comparison of pearson, spearman, transformation, and resampling approaches. *Psychological methods*, 17(3):399, 2012.
- [7] A de P Braga, APLF Carvalho, and Teresa B Ludermir. Redes neurais artificiais. *Teoria e Aplicações*, 2000.
- [8] Maritza Cabrera, Jason Leake, José Naranjo-Torres, Nereida Valero, Julio C Cabrera, and Alfonso J Rodríguez-Morales. Dengue prediction in latin america using machine learning and the one health perspective: a literature review. *Tropical Medicine and Infectious Disease*, 7(10):322, 2022.
- [9] Natalia Bruna Dias Campos et al. Vinte e dois anos de dengue em espaço urbano: estudo epidemiológico em belo horizonte, 1996-2017. 2017.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Frederico Coelho, Antonio P Braga, and Michel Verleysen. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *International Journal of Computational Intelligence Systems*, 9(4):726–733, 2016.
- [14] Frederico Gualberto Ferreira Coelho. *Semi-supervised Feature Selection*. PhD thesis, Universidade Federal de Minas Gerais, Université catholique de Louvain, 2013.
- [15] Departamento de Controle do Espaço Aéreo (DECEA). Bnd-met api. <https://api-bndmet.decea.mil.br/docs#/>, 2024. Acessado: 19 de Dezembro de 2024.
- [16] Instituto Nacional de Meteorologia (INMET). Banco de dados meteorológicos para ensino e pesquisa (bdmep). <https://bdmep.inmet.gov.br/#>, 2024. Acessado: 19 de Dezembro de 2024.
- [17] Instituto Nacional de Meteorologia (INMET). Dados meteorológicos históricos do inmet. <https://tportal.inmet.gov.br/dadoshistoricos>, 2024. Acessado: 19 de Dezembro de 2024.
- [18] Chuda Prasad Dhakal. A naïve approach for comparing a forecast model. *International Journal of Thesis Projects and Dissertations*, 5(1):1–3, 2017.
- [19] 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.
- [20] 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.

- [21] John T Hancock and Taghi M Khoshgoftaar. Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):94, 2020.
- [22] Instituto Nacional de Meteorologia. Mapas meteorológicos do brasil. <https://mapas.inmet.gov.br/>. Acesso em: 19 de dezembro de 2024.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Clarisse 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [32] 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.
- [33] José Eduardo Marques Pessanha. Onde está wally? ou onde se esconde o aedes aegypti. *Boletim Epidemiológico*, X(4):26, 2007.
- [34] 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.

- [35] Prefeitura de Belo Horizonte (PBH). Núcleo de pesquisa. Acesso em: 01 junho 2024.
- [36] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrej Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.
- [37] R Core Team. *R: A Language and Environment for Statistical Computing*, 2024. Acessado: 30 de Outubro de 2024.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] Olivia Lang Schultes, Maria Helena Franco Morais, 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.
- [44] Daniel Servén and Charles Brummitt. Pygam: Generalized additive models in python. <https://github.com/dswah/pyGAM>, 2018. Acessado: 10 de novembro de 2024.
- [45] 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.
- [46] 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.
- [47] Sediyama GC. Vianello RL, Pessanha JEM. Previsão de ocorrência dos mosquitos da dengue em belo horizonte com base em dados meteorológicos. 2006.
- [48] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A.

- Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [49] Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.
- [50] 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.