

UNIVERSIDADE FEDERAL DO PARANÁ

WAGNER HUGO BONAT
HENRIQUE SILVA DALLAZUANNA

INVESTIGANDO FATORES ASSOCIADOS A CONTAGENS
DE OVOS DE *Aedes Aegypti* COLETADOS EM
OVITRAMPAS EM RECIFE/PE

CURITIBA
2008

WAGNER HUGO BONAT
HENRIQUE SILVA DALLAZUANNA

INVESTIGANDO FATORES ASSOCIADOS A CONTAGENS
DE OVOS DE *Aedes Aegypti* COLETADOS EM
OVITRAMPAS EM RECIFE/PE

Trabalho de conclusão apresentado ao curso
de Bacharelado em Estatística, do Setor de
Ciências Exatas da Universidade Federal do
Paraná.

Orientador: Prof. Paulo Justiniano Ribeiro
Júnior

CURITIBA
2008

Agradecimentos

Gostaríamos de agradecer a Deus por nos permitir realizar este trabalho. Agradecemos ao Professor Paulo Justiniano Ribeiro Júnior pela orientação e confiança durante a realização deste trabalho.

Devemos agradecimentos também às equipes de campo e laboratório da Secretaria de Saúde do Recife/CVA e CPqAM/Fiocruz pelo intenso trabalho que produziu os dados aqui utilizados. Agradecemos também à toda equipe do projeto SAUDAVEL (<http://saudavel.dpi.inpe.br>), em particular ao Coordenador Dr Antônio Miguel Vieira Monteiro e à equipe do SAUDAVEL-Recife coordenada pela Dra Lêda Régis, a Wayner Vieira de Souza e a José Constantino Silveira pela troca de informações e por disponibilizar os dados e recursos utilizados neste trabalho.

E finalmente não poderíamos deixar de agradecer aos nossos pais, da parte de Henrique Silva Dallazuanna a sua mãe Maria de Fátima Silva Dallazuanna, da parte de Wagner Hugo Bonat a seus pais Hugo Bonat Neto e Leoni Chaste Bonat, pelo intenso apoio e dedicação durante os anos de graduação.

Este trabalho foi realizado com o apoio e recursos do LEG/UFPR.

Sumário

1	Introdução	1
2	Contexto e material	7
2.1	Área de estudo, Instrumentos e Técnicas de Campo	7
2.2	Construção das covariáveis	9
2.2.1	Covariáveis da armadilha	10
2.2.2	Covariáveis climáticas	11
3	Metodologia	13
3.1	Análise Fatorial	13
3.2	Modelo para dados de contagem	15
3.2.1	Representando uma função suave: regressão via <i>splines</i>	18
3.2.2	Bases de splines cúbicos	19
3.2.3	Controlando o grau de suavização com <i>penalized regression splines</i>	20
3.2.4	<i>A thin plate splines</i>	23
3.2.5	Escolhendo o parâmetro de suavização: validação cruzada	24
3.3	Modelos Aditivos	26
3.3.1	Representação de um modelo aditivo usando regressão <i>spline</i>	27
3.3.2	Ajustando um modelo aditivo por Mínimos Quadrados Penalizados	28
3.4	Modelos Aditivos Generalizados	28
4	Resultados	31
4.1	Análise descritiva e construção dos fatores	31
4.2	Aplicação do modelo estatístico	43
4.2.1	Brasília Teimosa	49
4.2.2	Casa Forte / Parnamirin	53
4.2.3	Dois Irmãos	57
4.2.4	Morro da Conceição	61

4.2.5	Engenho do Meio	64
5	Conclusão	68
6	Um olhar crítico aos dados.	71
7	Referências	75
8	Anexo	78

Lista de Tabelas

1	Descrição do experimento por bairros.	31
2	Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Brasília Teimosa.	35
3	Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Casa Forte / Parnamirin.	36
4	Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Dois Irmãos.	37
5	Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Morro da Conceição.	38
6	Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Engenho do Meio.	39
7	Composição dos fatores de Temperatura em percentuais por bairro.	41
8	Composição dos fatores de Umidade em percentuais por bairro.	42
9	Estimativa e p -valor do modelo combinado para todos os bairros.	45
10	Comparação entre os modelos ajustados através do Critério de <i>Akaike</i>	48
11	Ajuste do Modelo Aditivo Generalizado para o bairro Brasília Teimosa.	49
12	Ajuste do Modelo Aditivo Generalizado para o bairro Casa Forte / Parnamirin.	53
13	Ajuste do Modelo Aditivo Generalizado para o bairro Dois Irmãos.	57
14	Ajuste do Modelo Aditivo Generalizado para o bairro Morro Conceição.	61
15	Ajuste do modelo aditivo generalizado para o bairro Engenho do Meio.	64

Lista de Figuras

1	Mosquito <i>Aedes aegypti</i>	1
2	Formato geral de análise.	8
3	Histogramas e densidades estimadas em escala logarítmica (esquerda) e original (direita) do número de ovos nos bairros do estudo.	33
4	Superfície estimada da ocorrência de ovos de <i>Aedes aegypti</i> - Brasília Teimosa.	50
5	Ajuste do efeito temporal - Brasília Teimosa.	51
6	Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Brasília Teimosa.	52
7	Superfície estimada da ocorrência de ovos de <i>Aedes aegypti</i> - Casa Forte / Parnamirin.	54
8	Ajuste do efeito temporal - Casa Forte / Parnamirin.	55
9	Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Casa Forte / Parnamirin.	56
10	Superfície estimada da ocorrência de ovos de <i>Aedes aegypti</i> - Dois Irmãos.	58
11	Ajuste do efeito temporal - Dois Irmãos.	59
12	Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Dois Irmãos.	60
13	Superfície estimada da ocorrência de ovos de <i>Aedes aegypti</i> - Morro da Conceição.	62
14	Ajuste do efeito temporal - Morro da Conceição.	62
15	Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Morro da Conceição.	63
16	Superfície estimada da ocorrência de ovos de <i>Aedes aegypti</i> - Engenho do Meio.	66
17	Ajuste do efeito temporal - Engenho do Meio.	66
18	Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Engenho do Meio.	67

19	Observações da armadilha 09 do grupo 3, bairro Engenho do Meio.	72
20	Observações da armadilha 04 do grupo 1, bairro Casa Forte/Parnamirin	73

Investigando fatores associados a contagens de ovos de *Aedes aegypti* coletados em ovitrampas em Recife/PE

Resumo

O *Aedes aegypti* é o vetor da dengue, doença que pode resultar em epidemias. Estudos entomológicos são importantes, pois ajudam a entender a dinâmica de ploriferação do mosquito. Este trabalho tem como objetivo identificar fatores associados a contagens de ovos de *Aedes aegypti* a partir de dados coletados em um experimento de campo conduzido em bairros de Recife/PE. Os principais resultados mostram que plantas em vasos, charco/poça, garrafas, fossa externa, piscina, poço elevador laje sem telhado ou calhas são características de risco. As análises apontam para a preferência do vetor por residências, principalmente onde o abastecimento de água não é constante. A partir de covariáveis climáticas, temperatura e precipitação, utilizadas na análise, pode-se verificar que um período de altas temperaturas com um período de aproximadamente um mês para a reprodução, seguido de um período de alta umidade estão associados a elevadas contagens de ovos, sugerindo condições preferenciais para a elevação no número de ovos do mosquito.

As características espaço-temporais do experimento foram levadas em consideração na modelagem através de funções suaves das coordenadas geográficas das armadilhas e das datas de observações. Nessa abordagem utilizou-se a metodologia de modelos aditivos generalizados, que permitiu, através do termo espacial do modelo, a identificação de áreas onde as contagens de ovos são elevadas. O termo temporal permitiu identificar os meses do ano de maior abundância, observar efeitos sazonais e a tendência da série.

Palavras-Chave - Estudos entomológicos, *Aedes aegypti*, modelos aditivos generalizados, dengue, modelos espaço-temporais.

1 Introdução

A degradação do meio ambiente e os aspectos sócio-culturais afetam o cenário epidemiológico brasileiro, com impacto na mídia nacional e internacional, em virtude de epidemias de dengue, leptospirose, a recorrência da tuberculose, entre outras. Diante disso constatou-se a importância de modelar e identificar fatores de risco e proteção nas situações endêmicas e epidêmicas.

Neste sentido foi elaborado o Projeto SAUDAVEL¹, o qual pretende contribuir para aumentar a capacidade do setor de saúde no controle de doenças transmissíveis, demonstrando ser necessário desenvolver novos instrumentos para a prática da vigilância epidemiológica, incorporando aspectos ambientais, identificadores de risco e métodos automáticos e semi-automáticos, que permitam a detecção de surtos e seu acompanhamento no espaço e no tempo [Monteiro *et al.*, 2006].

Dentre as inúmeras doenças transmissíveis conhecidas nos dias de hoje, a dengue é uma das mais importantes e é atualmente um problema de saúde pública. O causador da dengue é um vírus, mas seus transmissores, são mosquitos do gênero *Aedes*, popularmente conhecidos como mosquito da dengue. Nos países de clima tropical, as condições do meio-ambiente favorecem o desenvolvimento e a proliferação do mosquito *Aedes aegypti* [Tauil, 2002], seu principal transmissor.

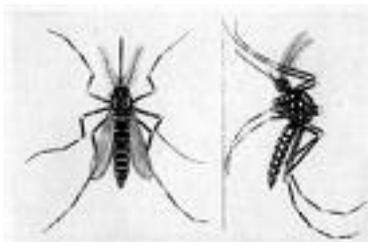


Figura 1: Mosquito *Aedes aegypti*.

O *Aedes aegypti* (figura 1), o mais comum do gênero *Aedes*, é um mosquito domesticado e urbano que se alimenta durante o dia. A transmissão da

¹Sistema de Apoio Unificado para a Detecção e Acompanhamento em Vigilância Epidemiológica (<http://saudavel.dpi.inpe.br>)

doença ocorre a partir da picada das fêmeas adultas do mosquito, pois somente elas são hematófogas. De 8 a 10 dias (período de incubação do vírus no mosquito) após ter sugado o sangue de uma pessoa contaminada, a fêmea infectada está apta a transmitir o vírus a indivíduos suscetíveis [Yang & Thomé, 2007].

No Brasil, a partir da década de 1980, iniciou-se um processo de intensa circulação do vírus da dengue, com epidemias explosivas que têm atingido todas as regiões brasileiras [Braga & Valle, 2007]. Em vista disso também a partir desta década diversas metodologias para a vigilância do vetor vêm sendo desenvolvidas e utilizadas no país.

O monitoramento de fatores de risco biológico relacionados aos vetores de doenças tem como finalidade o mapeamento de áreas de risco em determinados territórios. Para tanto, ele se utiliza da vigilância entomológica (presença do vetor, índices de infestação, características biológicas, tais como susceptibilidade aos inseticidas e aos vírus, avaliação da eficácia dos métodos de controle, entre outros), das ações de controle químico, biológico ou físico, e das relações com a vigilância epidemiológica quanto à incidência e prevalência da doença e ao impacto das ações realizadas.

Nos programas de controle de dengue, a vigilância entomológica é feita, principalmente, a partir de coletas de larvas, de acordo com proposta de [Connor & Monroe, 1923], para medir a densidade de *Aedes aegypti* em áreas urbanas. Essa metodologia consiste em vistoriar os depósitos de água e outros recipientes localizados nas residências e demais imóveis, como borracharias, ferros-velhos, cemitérios, entre outros tipos de imóveis considerados estratégicos, por produzirem grande quantidade de mosquitos adultos, para o cálculo dos índices de infestação predial (IIP) e de Breteal (IB), para mais detalhes ver [FNS, 2001] e [OPS, 1995].

Segundo [Braga & Valle, 2007] a coleta de larvas (ou pesquisa larvária, como é comumente chamada no Brasil) é importante para verificar o impacto das estratégias básicas de controle da doença, dirigidas a eliminação das larvas do vetor. Entretanto, não é um bom indicador para se medir a abundância do adulto, ineficaz para estimar o risco de transmissão, embora venha sendo usado com essa finalidade, como mostram as referências das

próprias autoras, ver [Nelson, 1995] e [Focks, 2000].

Outra metodologia adotada é a coleta de mosquitos adultos, cuja operacionalização para a estimativa do risco de transmissão é custosa e demorada. Ainda segundo [Braga & Valle, 2007], a correlação entre o número de vetores coletados (mosquitos adultos) e o número de humanos na área de coleta, que poderia fornecer o número de vetores adultos por pessoa, não é suficiente para quantificar o risco. Contudo, essa correlação se aproxima mais da realidade que os índices larvários, sobre isto ver [Focks, 2000].

Apesar disso, para avaliar a densidade do vetor, instalam-se armadilhas de oviposição e armadilhas para coleta de larvas, que visam estimar a atividade de postura. A armadilha de oviposição, também conhecida no Brasil como 'ovitrapa', é destinada à coleta de ovos. Em um recipiente de cor escura, adere-se um material àspero que permite a fixação dos ovos depositados. Em 1965, iniciou-se o uso de ovitrapas para a vigilância das populações adultas de *Aedes aegypti* [Fay & Perry, 1965]. Posteriormente, ficou demonstrada a superioridade dessas armadilhas em relação a pesquisa larvária para a verificação da ocorrência do vetor [Fay & Eliason, 1966].

As ovitrapas fornecem dados úteis sobre a distribuição espacial e temporal (sazonal). Dados obtidos com essa metodologia também são usados para verificar o impacto de vários tipos de medidas de controle que envolvem a redução do vetor com inseticidas. Segundo [Braga & Valle, 2007] embora as ovitrapas sejam muito úteis para a verificação da presença e distribuição de fêmeas do vetor, não devem ser usadas como ferramentas únicas para a estimativa do risco de dengue. O presente estudo baseia-se em dados de contagem de ovos coletados em ovitrapas conforme descrito em mais detalhes da Seção 2.

Diversas medidas para medir a infestação de *Aedes aegypti* em uma determinada área já foram propostas na literatura, a maioria tendo como base o uso das ovitrapas para capturar formas imaturas do mosquito. Em geral, a classe imatura é dividida pelos ecologistas em três níveis: ovo, larva e pupa. Na mensuração da densidade vetorial, cada método tem sua particularidade mas, todos lançam mão de amostras, nem sempre representativas, da população estudada. Um estudo bastante completo sobre medidas dos níveis de

infestação urbana para *Aedes* é encontrado em [Gomes, 2002].

Apesar do uso intenso das ovitrampas em estudos entomológicos, poucos foram os trabalhos encontrados que buscam relacionar a ocorrência das formas imaturas com fatores relacionados a própria condição da armadilha como, presença de recipientes potenciais de armazenamento de água, como tanques, caixa d'água, cisternas, além de fatores amplamente preconizados como vasos de plantas, presença de pneus entre outros.

O trabalho de [Santos, 1999] estudou os fatores associados à ocorrência de formas imaturas de *Aedes aegypti* na Ilha do Governador, no estado do Rio de Janeiro, contudo em um contexto diferente, sem o uso de ovitrampas e um delineamento amostral propriamente dito, o estudo foi realizado a partir dos dados da Fundação Nacional de Saúde (FNS). Os principais resultados obtidos pelo autor mostram que 58,04% dos criadores inspecionados foram constituídos por suportes para vasos com plantas, vasilhames de plástico ou vidro abandonados no peridomicílio. Maiores percentuais de criadouros positivos foram observados para pneus (1,41%), tanques, poços e cisternas (0,93%), e barris, tóneis e tinas (0,64%). Maiores percentuais de criadouros positivos durante o verão foram encontrados em grandes reservatórios de água e a dos criadores provenientes de lixo doméstico. No inverno foi verificado maior ocorrência em pequenos reservatórios de água para o uso doméstico. O autor também utilizou um modelo fatorial que mostrou como fator principal para determinar a ocorrência de fases imaturas de *Aedes aegypti* como sendo aquele que leva em consideração os fatores climáticos, as medidas utilizadas foram a média da pressão atmosférica, a temperatura máxima, temperatura mínima e a umidade relativa do ar.

O trabalho de [Ferreira & Chiaravalloti, 2007] buscou associar os índices de infestação larvária por *Aedes aegypti* com fatores socioeconômicos, em São José do Rio Preto, São Paulo. Os autores agruparam os setores censitários urbanos da cidade em 4 grupos segundo variáveis socioeconômicas, utilizando análise de componentes principais, e um quinto agrupamento com nível socioeconômico inferior aos demais não pertencente aos setores censitários urbanos. Calcularam-se, para cada área, os índices de Breteal (IB), predial (IP), de recipientes (IR), e as médias de recipientes existentes e pesquisados

por casa. Os valores dos índices de infestação não apresentaram diferenças significativas entre as áreas socioeconômicas 1 a 4, mais foram menores que para a quinta área. Os índices larvários não mostraram associação com os diferentes níveis socioeconômicos da área correspondente aos setores censitários urbanos, porém, os loteamentos da área 5 apresentaram os maiores valores desses indicadores.

Em estudo sobre a variação sazonal de *Aedes aegypti* [Serpa *et al.*, 2007], em Potin, São Paulo, os meses de julho, novembro e dezembro foram os de menor abundância do mosquito, também verificou-se a existência de relação significativa da ocorrência de formas imaturas com a temperatura máxima, porém não se encontrou relação com os fatores climáticos, temperatura mínima, pluviosidade e umidade relativa do ar.

Atualmente o cenário epidemiológico brasileiro reforça que estudos entomológicos devem ser feitos para buscar um melhor entendimento sobre os fatores que influenciam no comportamento reprodutivo do vetor, indicando assim fatores de risco e proteção para se evitar a proliferação do vetor e consequente queda nos indicadores de incidência e prevalência de doenças como a dengue.

Neste sentido, o objetivo deste estudo é determinar fatores de risco e proteção associados à ocorrência de ovos do mosquito *Aedes aegypti*, com base em dados de um experimento conduzido pelo "Projeto SAUDAVEL" na cidade de Recife/PE. Entende-se aqui como fatores de risco/proteção tanto covariáveis associadas à armadilha, como presença de recipientes grandes ou pequenos que possam conter água nas proximidades da armadilha, como também aspectos abióticos (climáticos) como temperatura, precipitação e umidade. Além de possíveis relações espaciais entre as armadilhas e também a possibilidade de uma relação temporal entre as coletas. Busca-se um modelo que leve em consideração a complexidade do ambiente em que o experimento foi desenvolvido e que possa ser usado tanto para entender a dinâmica de proliferação do mosquito, mostrando suas preferências para a reprodução, como também um modelo preditivo buscando evitar surtos entomológicos nas áreas em estudo.

Padrões espaciais eventualmente detectados pelo modelo podem destacar

zonas de persistência de alta infestação. Os padrões temporais podem auxiliar na compreensão da dinâmica temporal da população de mosquitos. Efeitos dos fatores do ambiente local dos pontos de coleta são avaliados buscando-se a seleção dos que apresentam relação com as contagens de ovos. Buscou-se ainda detectar condições climáticas refletidas em dados de temperatura e umidade em períodos que antecedem as coletas que possam criar condições para aumentos na contagem de ovos, bem como a determinação de tais períodos de defasagem de relevância.

O presente trabalho está dividido em cinco seções, esta primeira enfatiza a condição epidemiológica e entomológica dando uma visão ampla da situação brasileira, citando vários artigos que tratam sobre o mesmo assunto e também apresenta os objetivos gerais do trabalho. A Seção 2 visa descrever em detalhes o delineamento e forma de condução do experimento, descreve as variáveis utilizadas para a construção dos modelos e também sua forma de mensuração pela equipe de campo, além dos aspectos computacionais envolvidos na análise do experimento.

A Seção três traz um breve resumo sobre as metodologias estatísticas utilizadas, dando ênfase à análise fatorial e aos modelos aditivos generalizados. A Seção 4 traz os principais resultados da aplicação dos modelos estatísticos. A quinta e última seção traz uma breve discussão sobre os resultados e as principais conclusões e limitações do estudo, bem como aponta algumas possibilidades para investigações futuras.

2 Contexto e material

Nesta seção serão descritos o experimento, métodos de coleta de dados, e a construção das covariáveis que serão utilizadas posteriormente na composição dos modelos estatísticos.

2.1 Área de estudo, Instrumentos e Técnicas de Campo

O experimento está sendo desenvolvido na cidade de Recife/PE onde foram criteriosamente instaladas 564 armadilhas (ovitrampas) para o mosquito *Aedes aegypti* cuja a fêmea é o principal vetor da dengue. Estas armadilhas começaram a ser monitoradas em março de 2004. A cada 7 dias é feita a contagem de ovos encontrados em cerca de um quarto das armadilhas. Assim, em um ciclo de 28 dias, todas as armadilhas são monitoradas. O experimento está sendo realizado em 6 dos 94 bairros da cidade de Recife.

A rede de armadilhas foi instalada de modo a cobrir toda a extensão do bairro, caracterizando bem o tipo de delineamento utilizado para a coleta de dados. Até maio de 2007, período analisado aqui, foram realizadas 19.068 coletas, nas quais foram contados ao todo 14.829.557 ovos do mosquito. Cada armadilha contém uma lâmina (material áspero) na qual a fêmea do mosquito coloca os ovos, essas lâminas são recolhidas e a contagem dos ovos é feita em laboratório especializado. Os serviços de saúde locais e o laboratório de Entomologia são os coordenadores operacionais e logísticos, e responsáveis pela realização do experimento [Monteiro *et al.*, 2006]. Estes serviços também coletaram informações sobre as características dos locais das armadilhas que serão utilizadas nas análises.

Uma parte importante em projetos desta magnitude é a forma de armazenamento, tratamento e visualização dos dados, já que um experimento como este gera uma grande quantidade de dados que não são facilmente manipuláveis, requerendo-se para isto ferramentas específicas. Os dados coletados são inseridos em um banco de dados através de uma interface *Web* que foi desenvolvida pelo INPE-Instituto Nacional de Pesquisas Espaciais

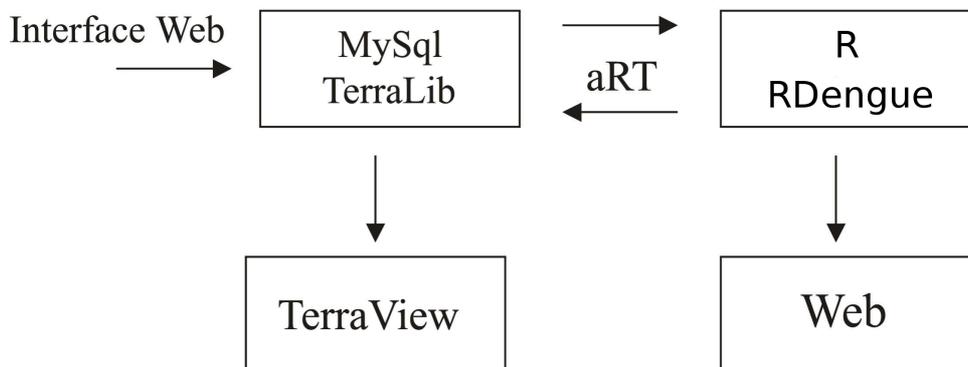


Figura 2: Formato geral de análise.

e CPqAM (Centro de pesquisas Aggeu Magalhães/Fiocruz, Recife). Esta interface foi projetada para evitar formas complexas de entrada de dados e fornecer algumas medidas espaço-temporais rápidas bem como permitir acesso e consultas de interesse para as análises.

O banco de dados do Recife SAUDAVEL, está implementado em **TerraLib**² tecnologia de código aberto *light-DBMS*, e *MySQL* Database Server³, como um repositório e sistema gerenciador de dados espaço-temporais, baseado no modelo espaço-temporal da **TerraLib** [Silveira *et al.*, 2004]. O banco de dados original do SAUDAVEL Recife fica no INPE e um backup é feito pelo LEG-Laboratório de Estatística e Geoinformação da UFPR.

O LEG tem particular importância para o experimento de Recife, já que, ele é o responsável por coordenar e implementar os modelos estatísticos. O LEG também é responsável por desenvolver tecnologias de integração entre a biblioteca **TerraLib** e um ambiente de computação e modelagem estatística, o **projeto R** [Development Core Team, 2007]. Neste sentido vem sendo desenvolvido o pacote **aRT** [Ribeiro *et al.*, 2007] e também o **RDengue** [Dallazuanna *et al.*, 2008].

A Figura 2 ilustra o formato geral da estrutura de dados e computacional relacionada ao projeto. Após a equipe de campo coletar as lâminas e estas serem analisadas e ovos contados pelo laboratório especializado, os dados

²<http://www.terralib.org>

³<http://www.mysql.com>

entram no banco através da interface *Web* e ficam armazenados no banco de dados geográficos, construído sobre a plataforma **TerraLib** com o *MySQL* Server Database. Isto feito, pode-se acessar o banco via **aRT**, assim todas as análises estatísticas são realizadas em um ambiente próprio no caso o **R**. Após as análises serem concluídas tem-se através do **aRT**, a opção de retornar os dados para o banco geográfico, podendo ser acessado por uma ferramenta de SIG (Sistema de Informação Geográfica) como é o caso do **TerraView**⁴, ou então gerar uma página *Web* para a visualização pública dos resultados. Como demonstração de tais análises, tem-se uma versão preliminar do pacote **RDengue** pode ser encontrada em <http://www.leg.ufpr.br/RDengue>.

O artigo de [Regis *et al.*, 2008] descreve de forma ampla o experimento SAUDAVEL/Recife, bem como todo o escopo do projeto que visa desenvolver metodologias e tecnologias para monitoramento de populações de *Aedes aegypti* através de contagens de ovos coletados em ovitrampas. Este trabalho se soma a este ambiente contribuindo com propostas de metodologias estatísticas para modelagem dos dados visando interpretações sobre a dinâmica da população de ovos do mosquito e a investigação dos fatores associados às contagens.

2.2 Construção das covariáveis

O banco do SAUDAVEL Recife proporcionou a construção de diversas covariáveis a serem utilizadas nos modelos para contagem de ovos. A seguir será descrita a forma de construção e os agrupamentos realizados para que as covariáveis pudessem ter uma interpretação prática bem como a possibilidade de comparação dentro de um modelo estatístico. O conjunto de covariáveis ainda foi dividido em dois grupos, o primeiro das covariáveis "locais" referentes à armadilha e o segundo de covariáveis "ambientais" referentes a fatores abióticos, representados por variáveis climáticas.

⁴<http://www.dpi.inpe.br/terraview/>

2.2.1 Covariáveis da armadilha

1. **Tipo de imóvel** - As armadilhas foram classificadas em duas categorias, residenciais(0) onde se considerou casas, apartamentos e residências com estabelecimentos comerciais. A segunda categoria denominada não residencial(1) considerou, estabelecimentos tais como: quartel, escolas, igrejas e outros.
2. **Presença de Quintal** - As armadilhas foram classificadas segundo a presença de quintal ou jardim. Caso apresentarem fazem parte da categoria (0) caso não são da categoria (1).
3. **Água ligada a rede geral** - As armadilhas foram classificadas em duas categorias segunda estarem ou não ligadas a rede geral de água. Caso estiverem fazem parte da categoria (0) caso não são da categoria (1).
4. **Frequência de abastecimento** - As armadilhas foram classificadas em duas categorias segundo a frequência de abastecimento de água. A categoria (0) é das armadilhas onde a frequência é diária e a categoria (1) caso não seja diária, podendo ser de dois em dois dias, três em três dias ou nenhum dia.
5. **Água canalizada no cômodo** - As armadilhas foram classificadas em duas categorias segunda a canalização de água em pelo menos um cômodo. A categoria (0) é das armadilhas que apresentam canalização de água em pelo menos um cômodo, a categoria (1) das armadilhas que não apresentam esta canalização.
6. **Presença de fatores de risco** - As armadilhas foram classificadas em duas categoriais. A categoria (0) é das armadilhas onde em sua proximidade encontram-se uma ou mais das seguintes características: plantas em vasos, charco/poça, garrafas, fossa externa, piscina, poço elevador, laje sem telhado ou calhas. A categoria (1) é das armadilhas que não tem em sua proximidade nenhuma destas características.

7. **Presença de recipientes grandes sem tampas** - As armadilhas foram classificadas em duas categoriais. A categoria (0) é das armadilhas onde em sua proximidade encontram-se uma ou mais das seguintes características: tanques, caixa d'água ou toneis sem tampa. A categoria (1) é das armadilhas que não tem em sua proximidade nenhuma destas características.
8. **Presença de recipientes grandes com tampas** - As armadilhas foram classificadas em duas categoriais. A categoria (0) é das armadilhas onde em sua proximidade encontram-se uma ou mais das seguintes características: tanques, caixa d'água ou toneis com tampa. A categoria (1) é das armadilhas que não tem em sua proximidade nenhuma destas características.
9. **Presença de recipientes pequenos sem tampas** - As armadilhas foram classificadas em duas categoriais. A categoria (0) é das armadilhas onde em sua proximidade encontram-se uma ou mais das seguintes características: jarros de barro ou baldes sem tampa. A categoria (1) é das armadilhas que não tem em sua proximidade nenhuma destas características.
10. **Presença de recipientes pequenos com tampas** - As armadilhas foram classificadas em duas categoriais. A categoria (0) é das armadilhas onde em sua proximidade encontram-se uma ou mais das seguintes características: jarros de barro ou baldes com tampa. A categoria (1) é das armadilhas que não tem em sua proximidade nenhuma destas características.

2.2.2 Covariáveis climáticas

As covariáveis climáticas disponíveis no banco de dados e utilizadas nas análises foram precipitação, umidade relativa do ar, temperatura máxima e mínima. A mensuração destas covariáveis é feita diariamente para cada um dos bairros em análise por cinco estações de monitoramento. Portanto,

para fins de modelagem estatística, são necessárias decisões e manipulações destes dados para serem relacionados com os de coleta de dados realizadas semanalmente.

Em um primeiro momento as observações foram agrupadas semanalmente, sendo que para as covariáveis precipitação e umidade relativa do ar, foi tirada a média dos sete dias da semana, para as variáveis temperatura máxima foi utilizada a maior temperatura da semana. Para a temperatura mínima a menor temperatura da semana.

Após isto foram contabilizadas todas as covariáveis com uma defasagem de até doze semanas da observação. Por exemplo, se uma observação foi realizada no dia 10 do mês de junho, as covariáveis climáticas foram contabilizadas na semana do dia 10 de junho e mais 11 semanas atrás.

Como era esperado e confirmado nas análises dos dados, dados climáticos de semanas sequenciais apresentavam um alto nível de correlação. Desta forma, foram criados fatores através da análise fatorial. As cargas fatoriais definiram os fatores que resumiram os dados, levando assim em consideração toda a informação disponível referente a estas covariáveis e evitando um problema conhecido como multicolinearidade. Tal procedimento também permitiu diminuir a quantidade de covariáveis a serem selecionadas para compôr o modelo.

3 Metodologia

Nesta seção serão apresentadas inicialmente as técnicas de análise fatorial utilizada para compôr os fatores climáticos. Na sequência será feita uma breve revisão de alguns modelos para dados de contagem, incluindo o ajuste de funções suaves através de modelos aditivos generalizados.

3.1 Análise Fatorial

A análise fatorial tem o propósito essencial de tentar descrever as relações de covariância entre diversas variáveis em função de poucas, não observáveis, quantidades aleatórias denominadas **fatores**. Sob o modelo de fatores cada variável é representada por uma função linear de uma pequena quantidade de fatores comuns, não observáveis, e de uma simples variável latente. Os fatores comuns geram as covariâncias entre as observações e os termos específicos contribuem somente para as variâncias de suas respostas relacionadas. Os coeficientes de fatores comuns não são restritos à condição de ortogonalidade, o que confere generalidade, apesar de se exigir normalidade dos dados e a determinação, a priori, do número de fatores.

O uso da análise fatorial neste trabalho foi uma forma considerada para acomodar em um modelo estatístico um grande número de covariáveis que são fortemente correlacionadas.

As covariáveis referentes ao clima (precipitação, umidade do ar, temperatura máxima e mínima) foram consideradas com uma defasagem de até doze semanas da observação. Com isso foi detectada uma estrutura de correlação entre as observações das semanas sequenciais. O uso das variáveis originais pode levar a resultados inconclusivos sobre a influência de cada covariável devido ao problema da multicolinearidade. Foi com o intuito de evitar este problema que a análise fatorial foi empregada. O método de estimação do modelo fatorial neste trabalho foi o de máxima verossimilhança, implementado no software **R** [Development Core Team, 2007].

Assumindo-se que os fatores comuns F e os fatores específicos ϵ possuem

distribuição normal, estimativas de máxima verossimilhança podem ser obtidas. A partir da suposição de normalidade, pode-se concluir que $\underline{X}_j - \underline{\mu} = L\underline{F}_j + \underline{\epsilon}_j$, também é normalmente distribuído e portanto a função de verossimilhança é dada por:

$$L(\underline{\mu}, \Sigma) = (2\Pi)^{-np/2} |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[\Sigma^{-1} \left(\sum_{j=1}^n (\underline{X}_j - \overline{\underline{X}})(\underline{X}_j - \overline{\underline{X}})^T + n(\overline{\underline{X}} - \underline{\mu})(\overline{\underline{X}} - \underline{\mu})^T \right) \right] \right\} \quad (1)$$

a qual depende de L e Ψ por meio de $\Sigma = LL^T + \Psi$.

Devido à multiplicidade de escolhas para L dadas por transformações ortogonais é imperativo uma restrição de unicidade computacional por:

$$L^T \Psi^{-1} L = \Delta \quad (2)$$

em que Δ é uma matrix diagonal.

Os estimadores de máxima verossimilhança \widehat{L} e $\widehat{\Psi}$ devem ser obtidos por maximização numérica de (1). A maximização de (1) sujeita à condição de unicidade 2 deve satisfazer:

$$(\widehat{\Psi}^{-1/2} S_n \widehat{\Psi}^{-1/2})(\widehat{\Psi}^{-1/2} \widehat{L}) = \widehat{\Psi}^{-1/2} \widehat{L}(I + \widehat{\Delta}) \quad (3)$$

[Lawley, 1942] mostra que o estimador de $\widehat{\Delta}$ é dado por:

$$\widehat{\Delta} = \widehat{L}^T \widehat{\Psi}^{-1} \widehat{L} \quad (4)$$

Em [Johnson & Wichern, 1998] pode ser encontrado uma descrição completa do processo iterativo utilizado para maximizar a função em (1) e consequentemente obter as estimativas de L e Ψ , e também as expressões para o cálculo das cargas fatoriais (\widehat{L}_z), das variâncias específicas ($\widehat{\Psi}_z$) e das comunalidades.

Neste trabalho o objetivo principal é contruir fatores que agrupem, re-

sumam e permitam a interpretação das covariáveis ambientais do modelo. O interesse principal é portanto nas cargas fatoriais, já que elas definem a qual fator determinada variável irá pertencer e qual a proporção do fator devido à ela.

3.2 Modelo para dados de contagem

O modelo log-linear Poisson é sem dúvida o modelo mais utilizado quando se trata de dados de contagem. Porém em alguns casos pode ocorrer o fenômeno da superdispersão, refletindo o fato de que a variância nas observações não reflete as propriedades do modelo de Poisson, sendo, em geral, maior do que a predita pelo modelo. Nesses casos a suposição de distribuição de Poisson para a resposta é inadequada sendo necessário o uso de modelos alternativos. Uma opção inicial seria aplicar um modelo de quase-verossimilhança com parâmetro extra de dispersão, que produz as mesmas estimativas do modelo Poisson porém corrigindo os erros padrão das estimativas [McCullagh & Nelder, 1989]. Contudo, neste trabalho, foi utilizada uma outra abordagem, assumindo-se a distribuição Binomial Negativa a qual, segundo [Paula, 2004], permite uma análise mais completa dos dados do que os modelos de quase-verossimilhança.

O fenômeno da superdispersão ocorre quando é esperada uma distribuição de Poisson para a resposta, porém a variância observado nos dados é maior que a resposta média. Uma causa provável desse fenômeno é a heterogeneidade das unidades amostrais, que pode ser devido a variabilidades inter-unidades experimentais. Este fato foi detectado no caso dos dados do experimento SAUDAVEL/Recife onde as unidades experimentais, dadas pelas armadilhas, são notadamente diferentes umas das outras. Do ponto de vista das propriedades probabilidade, isso pode ser formulado, introduzindo uma variável latente que denotaremos por Z e especificamos o modelo como se segue. Supondo um conjunto fixo $x = (x_1, \dots, x_p)^T$ de valores de p variáveis explicativas, $[Y|z]$ tem média z e variância z , no entanto Z , que é não observável, varia nas unidades amostrais com x fixo, de modo que $E(Z) = \mu$.

Então,

$$\begin{aligned} E(Y) &= E_z[E_y(Y|Z)] = E_z[Z] = \mu \quad \text{e} \\ \text{Var}(Y) &= E_z[\text{Var}_y(Y|Z)] + \text{Var}_z[E_y(Y|Z)] \\ &= \mu + \text{Var}(Z) \end{aligned}$$

E como $\text{Var}(Z)$ é uma quantidade não negativa, tal relação especifica um modelo de superdispersão. Note-se que este resultado não há suposições sobre distribuições de probabilidade para Y e Z . Adicionalmente, pode-se supor que $[Y|z]$ tem distribuição Poisson com média z e função de probabilidade denotada por $f(y|z)$ e que Z segue uma distribuição gama de média μ e parâmetro de dispersão $k = \phi\mu$ cuja função de densidade será denotada por $g(z; \mu, k)$. Logo, a distribuição marginal $[Y]$ tem função de probabilidade dada por

$$\begin{aligned} P(Y = y) &= \int_0^\infty f(y|z)g(z; \mu, k)dz \\ &= \frac{\Gamma(y+k)\phi^k}{\Gamma(y+1)\Gamma(k)(1+\phi)^{y+k}} \\ &= \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{\phi}{1+\phi}\right)^k \left(\frac{1}{1+\phi}\right)^y \\ &= \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} (1-\pi)^k \pi^y \end{aligned}$$

em que $\pi = \frac{1}{1+\phi}$, note que $\text{Var}(Z) = \frac{\mu^2}{k}$ de modo que $\text{Var}(Y) = \mu + \frac{\mu^2}{k} = \frac{\mu(1+\phi)}{\phi}$. Portanto, $[Y]$ tem distribuição binomial negativa.

Supondo que Y_1, \dots, Y_n são variáveis independentes de modo que $Y_i \sim BN(\mu_i, \phi)$, em que $E(Y_i) = \mu_i$ e $\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\phi}$ e parte sistemática dada por $g(\mu_i) = x_i^T \beta$ em que $g(\cdot)$ é uma função de ligação. Definindo $\theta = (\beta, \phi)^T$ o logaritmo da função de verossimilhança fica dado por

$$L(\theta) = \sum_{i=1}^n \left[\log \left\{ \frac{\Gamma(\phi + y_i)}{\Gamma(y_i + 1)\Gamma(\phi)} \right\} + \phi \log \phi + y_i \log y_i - (\phi + y_i) \log(\mu_i + \phi) \right]$$

em que $\mu_i = \exp(x_i^T \beta)$. As funções escore para β e ϕ ficam, respectivamente,

dadas por:

$$U_{\beta}(\theta) = X^T W F^{-1}(y - \mu)$$

e

$$U_{\phi}(\theta) = \sum_{i=1}^n \left[\Phi(\phi + y_i) - \Phi(\phi) - \frac{y_i + \phi}{\phi + \mu_i} + \log \phi / (\phi + \mu_i) + 1 \right]$$

em que X é a matriz do modelo com linhas $x_i^T, i = 1, \dots, n$, $W = \text{diag}(\omega_1, \dots, \omega_n)$ com $\omega_i = \frac{(d\mu_i/d\eta_i)^2}{(\mu_i^2 \phi^{-1} + \mu_i)}$, $F = \text{diag}(f_1, \dots, f_n)$ com $f_i = \frac{d\mu_i}{d\eta_i}$, $y = (y_1, \dots, y_n)^T$, $\mu = (\mu_1, \dots, \mu_n)^T$ e $\Phi(\cdot)$ é a função digama. As estimativas de máxima verossimilhança podem ser obtidas através de algoritmos iterativos como o método de Newton-Raphson. Para maiores detalhes sobre procedimentos e algoritmos de estimação ver [Paula, 2004]. Ainda assumindo que ϕ é fixo a função desvio (*deviance*) fica dada por

$$D^*(y; \hat{\mu}) = 2 \sum_{i=1}^n \left[\phi \log \frac{\hat{\mu}_i + \phi}{y_i + \phi} + y_i \log \frac{y_i(\hat{\mu}_i + \phi)}{\hat{\mu}_i(y_i + \phi)} \right]$$

em que $\hat{\mu}_i = g^{-1}(x_i^T, \hat{\beta})$ obtidos os estimadores de máxima verossimilhança de β . Sob a hipótese de que o modelo adotado está correto $D^*(y; \hat{\mu}_i)$ segue para ϕ grande e μ_i grande \forall_i uma qui-quadrado com $(n - p)$ graus de liberdade.

O modelo descrito, apesar de muito utilizado, tem uma restrição quanto ao relacionamento da variável resposta com as covariáveis, assumindo que tal relacionamento seja linear na escala da função de ligação, o que, em diversas situações não reflete a realidade. Em particular, no caso do presente estudo onde deseja-se incluir efeitos espaciais através das coordenadas geográficas das armadilhas, em uma forma flexível que descreva tal relação. É então necessária uma extensão do modelo anterior que comporte a inclusão de uma ou mais covariáveis sem assumir a priori que o relacionamento com a resposta seja linear.

Uma das possíveis abordagens é o modelo aditivo generalizado [Hastie & Tibishirani, 1990] que pode ser descrito como um extensão um modelo linear generalizado como o anterior, porém com um ou mais preditores lineares envolvendo a soma de funções suaves (*smooth functions*) das covariáveis. O modelo torna-se semi-paramétrico e pode ser escrito da seguinte

forma:

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots$$

onde

$$\mu_i \equiv E(Y_i) \quad \text{e} \quad [Y_i] \sim \{\text{família exponencial}\}.$$

Y_i é a variável resposta com distribuição de probabilidade na família exponencial, X_i^* é uma linha da matriz do modelo, para a parte estritamente paramétrica, θ é o vetor de parâmetros correspondentes e as f_j são funções suaves das covariáveis x_k . As funções suaves podem ter mais de uma covariável como argumento conforme ilustrado por f_3 nessa expressão. O modelo definido desta forma proporciona flexibilidade na especificação da forma da relação entre a variável resposta e as covariáveis. Casos particulares como a especificação do modelo apenas em termos de funções suaves também são possíveis, evitando desta forma qualquer suposição de relacionamento apenas linear entre as variáveis. Entretanto, a flexibilidade de tais modelos vem acompanhada de dois novos problemas teóricos:

1. como representar as funções suaves;
2. como estimar os parâmetros envolvidos neste modelo.

Estes dois tópicos são discutidos no restante desta subseção. Sem perda de generalidade, a representação de uma função suave é discutida considerando um modelo com função de ligação identidade, contendo apenas uma função suave de uma covariável que pode ser escrito na forma

$$y_i = f(x_i) + \epsilon_i \tag{5}$$

onde y_i é a variável resposta, x_i uma covariável, f é uma função suave e ϵ_i são v.a. normais, independentes e identicamente distribuídas com média zero e variância σ^2 .

3.2.1 Representando uma função suave: regressão via *splines*

Para estimar f utilizando métodos tradicionais como mínimos quadrados e máxima verossimilhança é necessário que f seja representada em (5) como

um modelo linear. Isto pode ser feito escolhendo-se uma função de base e definindo o espaço para as funções de qual f é um elemento.

A função de base escolhida deve ser conhecida e totalmente tratável analiticamente. Se b_j é o j -ésimo termo de tal função de base, assume-se a seguinte representação:

$$f(x) = \sum_{j=1}^q b_j(x)\beta_j; \quad (6)$$

dados alguns parâmetros β_j desconhecidos. Substituindo (6) em (5) tem-se claramente um modelo linear.

Um exemplo simples de função de base pode ser dado pelas bases polinomiais. Neste caso supõe-se que f é dada por um polinômio de quarta ordem, então f contém todos os polinômios de ordem menor. Uma base para este espaço é $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, b_4(x) = x^3$ e $b_5(x) = x^4$. Substituindo-se em (6) tem-se:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5$$

e substituindo agora em (5) chega-se a um modelo linear

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i. \quad (7)$$

Bases polinomiais tendem a ser muito úteis em situações em que o foco de análise são as propriedades de f na vizinhança de um único ponto. Porém, quando o interesse é em toda a extensão do domínio da função as bases polinomiais tem alguns problemas. Uma base *spline* tem melhor performance perante uma grande quantidade de situações e pode ser mostrado que elas tem boas propriedades teóricas [Wood, 2006].

3.2.2 Bases de splines cúbicos

Uma função univariada pode ser representada usando uma *spline* cúbica. Uma *spline* cúbica é uma curva composta por seções de polinomiais cúbicas juntas de modo que juntos componham uma função contínua que permita primeira e segunda derivada. Os pontos onde as seções se juntam são co-

nhecidos como *knots* do *spline*. Para um *spline* comum os nós (*knots*) ocorrem onde quer que haja uma curva de referência. Para regressão *spline* que é o interesse aqui, a localização dos nós deve ser especificada.

Tipicamente os *knots* são espalhados de maneira uniforme em toda a extensão dos valores observados de x , ou nos quantis da distribuição de x . Seja qual for o método para encontrar os *knots* eles serão denotados por $x_i^* : i = 1, \dots, q - 2$.

Dada a localização dos *knots*, tem-se muitas formas alternativas e equivalentes de escrever a base para a *spline* cúbica. Uma base simples para ser utilizada, pode ser encontrada nos livros de [Wahba, 2000] e [Gu, 2002]. Embora a expressão desta função de base dada a seguir seja ligeiramente intimidadora e a definição dada seja um pouco vaga, seu uso na prática é fácil.

$$R(x, z) = \frac{[(z - \frac{1}{2})^2 - \frac{1}{12}][(x - \frac{1}{2})^2 - \frac{1}{12}]}{4} - \frac{[(|x - z| - \frac{1}{2})^4 - \frac{1}{2}(|x - z| - \frac{1}{2})^2 + \frac{7}{240}]}{24}$$

Para mais detalhes sobre esta função de base [Gu, 2002, p.37]. Usar esta base de *spline* cúbica para f significa que o modelo em (5) torna-se um modelo linear da forma $y = X\beta + \epsilon$, onde a i -ésima linha da matriz do modelo é:

$$X_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), \dots, R(x_i, x_{q-2}^*)]$$

Desta forma o modelo pode ser estimado por mínimos quadrados ordinários que é simples computacionalmente. Apesar deste modelo ser satisfatório, a escolha do grau de suavidade do modelo é essencialmente arbitrária, controlada pela dimensão q da base escolhida.

3.2.3 Controlando o grau de suavização com *penalized regression splines*

Uma forma simples para escolher o grau de suavização, é tentar fazer uso de testes de hipóteses, para selecionar q por uma seleção da forma *backward*. Porém tal opção é um tanto problemática, dado que um modelo com $k - 1$ *knots* espalhados de forma uniforme geralmente não é aninhado a um modelo

com k *knots* espalhados de forma uniforme. Outra opção seria começar com uma grade de valores para os *knots* e dar saltos sequenciais, como parte da seleção *backward*. Entretanto, o resultado de espaçamentos diferentes para os *knots* conduzem a modelos com baixo desempenho, além de os resultados ficarem muito dependentes da posição escolhida para os *knots*.

Uma alternativa para controlar a suavização sem alterar a dimensão da base é manter a dimensão da base fixa, em um tamanho um pouco maior do que se acha necessário, porém o controle da suavização do modelo é feito adicionando uma penalização à "rugosidade" -- falta de suavidade (*wigglines*) na função objetivo do ajuste por mínimos quadrados. Por exemplo, o ajuste usual feito via mínimos quadrados visa minimizar a seguinte função objetivo:

$$\|y - X\beta\|^2.$$

enquanto que, seguindo as idéias de penalização, minimiza-se a seguinte função objetivo:

$$\|y - X\beta\|^2 + \lambda \int_0^1 [f''(x)]^2 dx$$

em que a integral do quadrado da segunda derivada do modelo penalizado é o termo *wiggly* de penalização à falta de suavidade. O relacionamento entre o modelo ajustado e o modelo suavizado é controlado pelo parâmetro de suavização λ . Tem-se desta forma que $\lambda \rightarrow \infty$ vai alisar muito os dados. Quando $\lambda = 0$ resulta em não penalização e uma simples regressão é ajustada.

Como f é uma função linear nos parâmetros β_i , a penalização pode sempre ser escrita como uma forma quadrática em relação a β .

$$\int_0^1 [f''(x)]^2 dx = \beta^T S \beta$$

onde S é uma matriz de coeficientes conhecidos. É neste momento que o uso de uma base *spline* mostra suas vantagens, já que, $S_{i+2,j+2} = R(x_i^*, x_j^*)$ para $i, j = 1, \dots, q - 2$ quando a primeira e segunda linhas e colunas são iguais a zero. Desta forma, o problema a ser resolvido pela regressão por *splines* é

simplesmente minimizar:

$$\|y - X\beta\|^2 + \lambda\beta^T S\beta \quad (8)$$

em relação a β . O problema de estimar o grau de suavização para o modelo é agora um problema de estimação do parâmetro λ . Porém, antes de proceder-se a estimação de λ considere a estimação de β dado que λ é assumido como conhecido. É razoavelmente simples mostrar que a equação que minimiza (8), é o estimador de mínimos quadrados penalizados para β isto é;

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T y$$

Similarmente ao ajuste de mínimos quadrados ordinários, a matriz de influência ou matriz chapéu, A para o modelo pode ser escrita como:

$$A = X(X^T X + \lambda S)^{-1} X^T,$$

lembrando que $\hat{\mu} = Ay$. Na prática esta expressão não é a usada computacionalmente, porque é mais útil utilizar a estabilidade numérica dada por matrizes ortogonais. Nos cálculos computacionais a equação utilizada é:

$$\left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda B} \end{bmatrix} \beta \right\|^2 = \|y - X\beta\|^2 + \lambda\beta^T S\beta$$

em que B é a matriz raiz quadrada da matriz S tal que $B^T B = S$. A soma de termos ao quadrado, na esquerda é equivalente a soma de quadrados da função objetivo para um modelo em que a matriz foi aumentada por uma matriz raiz quadrada de penalidades, quando o vetor de dados da resposta for aumentada por q zeros. B pode ser obtida facilmente por decomposição espectral ou decomposição de Choleski, e uma vez obtida, o problema de mínimos quadrados pode ser resolvido utilizando métodos ortogonais.

Desta forma, nem a escolha exata de q nem a escolha da posição exata dos *knots* tem muita influência no modelo ajustado. A escolha de λ é agora o ponto crucial para determinar a flexibilidade do modelo e a forma da função

\hat{f} .

De forma geral o que precisamos até aqui para estimar o modelo, é escolher a dimensão da base q , a localização dos *knots* e o valor para o parâmetro de suavização. Como dito acima as duas primeiras escolhas não vão influenciar de forma significativa no modelo ajustado isso pode ser visto também em [Wood, 2006]. A escolha de crucial importância neste momento é a do valor do parâmetro de suavização λ .

3.2.4 A *thin plate splines*

Uma outra opção de função *spline* é a *thin plate*. Segundo [Wood, 2006] é uma solução elegante e geral para o problema de estimar uma função suave de variáveis preditoras múltiplas. [Wahba, 2000] mostra que *thin plate splines* são uma generalização natural da *spline* polinomial univariada para duas ou mais dimensões via uma generalização do problema variacional em

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_1(i), x_2(i)))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx.$$

O problema variacional no espaço duplo Euclidiano (\mathbf{R}^2) é encontrar f em um espaço X apropriado para minimizar:

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_1(i), x_2(i)))^2 + \lambda \sum_{v=0}^m \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \binom{m}{v} \left(\frac{\partial f}{\partial x_1^v \partial x_2^{m-v}} \right)^2 dx_1 dx_2.$$

Note que os limites de integração são $\pm\infty$. Se um limite finito é especificado, então um problema do valor limite deve ser resolvido numericamente.

A dificuldade com *thin plate splines* é o custo computacional, dado que estes suavizadores têm tantos parâmetros desconhecidos quanto dados (estritamente, número de combinações únicas do preditor). Exceto no caso do preditor simples, o custo computacional da estimação do modelo é proporcional ao cubo do número de parâmetros.

Apesar disto neste trabalho ele foi preferido e amplamente usado exceto em casos em que o algoritmo não convergiu e nestes casos foi substituído

por *splines* cúbicos. Uma de suas principais características é a isotropia da penalidade das ondulações, onde tais ondulações são em todas as direções igualmente tratadas, com o ajuste inteiramente invariante para a rotação do sistema de coordenadas das covariáveis preditoras. Segundo [Wood, 2006, pg. 228] a *thin plate splines* é adequada para suavizar interações entre variáveis medidas na mesma unidade, como coordenadas geográficas, onde a isotropia é assumida como adequada.

3.2.5 Escolhendo o parâmetro de suavização: validação cruzada

Se λ for muito pequeno o modelo não alisará os dados adequadamente, e se o λ for muito grande o modelo alisará muito os dados. Em ambos os casos isto significa que a *spline* estimada \hat{f} não é uma boa aproximação para a função real f . A melhor escolha para λ é aquela que mais aproxima \hat{f} da real função f . Um critério adequado para escolher λ seria escolher o λ que minimiza a seguinte expressão:

$$M = \frac{1}{n} \sum_{i=1}^n (\hat{f}_i - f_i)^2,$$

onde a notação $\hat{f}_i \equiv \hat{f}(x_i)$ e $f_i \equiv f(x_i)$ foi adotada por conveniência.

Note que f é desconhecida, então M não pode ser usada diretamente. Entretanto é possível derivar uma estimativa da $E(M) + \sigma^2$ que é o erro quadrático médio esperado ao prever uma nova variável. Define-se $\hat{f}^{[-i]}$ como o modelo ajustado com todos os dados exceto y_i , e define-se o escore ordinário da validação cruzada como:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^n (f_i^{[-i]} - y_i)^2.$$

Este escore é o resultado de ajustar o modelo com uma parte dos dados, usar o modelo para prever os valores retirados, calcular a diferença entre os retirados e os preditos e, com estas diferenças, calcular uma diferença média,

passando por todos os dados. Substituindo $y_i - f_i + \epsilon_i$,

$$\begin{aligned}\nu_0 &= \frac{1}{n} \sum_{i=1}^n (\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{f}_i^{[-i]} - f_i)^2 - (\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2.\end{aligned}$$

Se $E(\epsilon_i) = 0$, e ϵ_i e $\hat{f}_i^{[-i]}$ são independentes, o segundo termo da soma desaparece se a esperança for dada por:

$$E(\nu_0) = \frac{1}{n} E \left(\sum_{i=1}^n (\hat{f}_i^{[-i]} - f_i)^2 \right) + \sigma^2.$$

Agora, $\hat{f}^{[-i]} \equiv \hat{f}$ com igualdade quando o tamanho da amostra tende a infinito. Assim $E(\nu_0) \equiv E(M) + \sigma^2$ com igualdade quando o tamanho da amostra tende a infinito. Sendo assim escolher λ que minimize ν_0 é uma aproximação razoável para minimizar M . Este processo de escolha de λ para minimizar ν_0 é conhecido como "validação cruzada ordinária". Este método é uma aproximação razoável, já que ele minimiza o erro quadrático de predição. Se os modelos forem julgados apenas por sua capacidade preditiva, modelos mais complicados serão sempre escolhidos ao invés de modelos mais simples. Desta forma, escolher um modelo que tem máxima sua habilidade em prever os dados fora dos dados observados não resolve o problema, além de ser computacionalmente ineficiente calcular ν_0 para todos os dados e ajustar o modelo para cada um dos n componentes do conjunto de dados. Felizmente isto pode ser contornado uma vez que pode ser mostrado que:

$$\nu_0 = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \hat{f}_i)^2}{(1 - A_{ii})^2}$$

onde \hat{f} é a estimativa vinda do ajuste com todos os dados e A corresponde a matriz de influência. Na prática os pesos $1 - A_{ii}$, são frequentemente substituídos pela média ponderada, $tr(I - A)/n$, chegando assim ao escore

da validação cruzada generalizada

$$\nu_g = \frac{n \sum_{i=1}^n (y_i - \hat{f}_i)^2}{[tr(I - A)]^2}.$$

O escore da validação cruzada generalizada (GCV) tem vantagens computacionais sobre o escore de validação cruzada ordinária (OCV), e também tem vantagens em termos de invariância ver [Wahba, 2000, pg. 53]. É também possível mostrar que GCV minimiza $E(M)$ quando a amostra tende a infinito ou é relativamente grande.

3.3 Modelos Aditivos

Suponha que se tem duas covariáveis, x e z , para explicar uma variável resposta y . Assume-se ainda que uma estrutura simples como a apresentada a seguir seja apropriada para este modelo,

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i. \quad (9)$$

Novamente as f_j são funções suaves e $\epsilon_i \stackrel{i.i.d}{\sim} N(0, \sigma^2)$. Há dois pontos a notar sobre este modelo. Primeiro, a suposição de efeitos aditivos é razoavelmente forte; $f_1(x) + f_2(z)$ é um caso especial restritivo para funções suaves de duas covariáveis sendo um caso especial de uma função ampla também de duas covariáveis $f(x, z)$. Segundo, o fato de que o modelo agora contém mais de uma função suave que deve ser introduzida e identificada. As funções f_1 e f_2 são estimadas com uma certa constante aditiva. Para perceber isto, note que toda constante pode ser simultaneamente adicionada a f_1 e subtraída de f_2 , sem mudar as predições do modelo.

Desde que as suposições acima sejam aceitas o modelo aditivo pode ser representado por regressão *spline* e estimado por mínimos quadrados penalizados com o grau de suavização pode ser selecionado por validação cruzada da mesma forma que para o modelo de uma covariável.

3.3.1 Representação de um modelo aditivo usando regressão *spline*

Cada função suave em (9) pode ser representada usando uma regressão *spline*. Usando a mesma base de *splines* vinda da Seção 3.2.2,

$$f_1(x) = \delta_1 + x\delta_2 + \sum_{j=1}^{q_1-2} R(x, x_j^*)\delta_{j+2}$$

e

$$f_2(z) = \gamma_1 + z\gamma_2 + \sum_{j=1}^{q_2-2} R(z, z_j^*)\gamma_{j+2}$$

onde δ_j e γ_j são parâmetros desconhecidos para f_1 e f_2 respectivamente, q_1 e q_2 são o número de parâmetros desconhecidos para f_1 e f_2 , enquanto que x_j^* e z_j^* são as localizações dos *knots* para as duas funções.

O principal problema em estimar um modelo aditivo é que δ_1 e γ_1 são confundidos. A maneira mais simples para tratar isto é condicionar um deles a zero, sendo assim diz-se que $\gamma_1 = 0$. Fazendo isto, é fácil ver que o modelo aditivo pode ser escrito na forma de um modelo linear, $y = X\beta + \epsilon$ onde a i -ésima linha da matriz do modelo é agora dada por

$$X_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), \dots, R(x_i, x_{q_1-2}^*), z_i, R(z_i, z_1^*), \dots, R(z_i, z_{q_2-2}^*)],$$

o vetor de parâmetros é $\beta = [\delta_1, \delta_2, \dots, \delta_{q_1}, \gamma_2, \gamma_3, \dots, \gamma_{q_2}]^T$.

O termo de rugosidade (*wiggleness*) de cada função pode ser mensurado exatamente como descrito na seção 3.2.3,

$$\int_0^1 f_1''(x)^2 dx = \beta^T S_1 \beta \quad \text{e} \quad \int_0^1 f_2''(x)^2 dx = \beta^T S_2 \beta$$

onde os elementos de S_1 e S_2 são zeros exceto onde $S_{1i+2, j+2} = R(x_i^*, x_j^*)$ para $i, j = 1, \dots, q_1 - 2$ e $S_{2i+q_1+1, j+q_1+1} = R(z_i^*, z_j^*)$ para $i, j = 1, \dots, q_2 - 2$. É possível utilizar um grande número de bases *splines* diferentes da descrita aqui, porém somente detalhes serão mudados e não o princípio geral, uma vez que a base foi escolhida, as matrizes e os coeficientes do modelo podem

ser obtidos imediatamente.

3.3.2 Ajustando um modelo aditivo por Mínimos Quadrados Penalizados

Os parâmetros β do modelo em (9) são obtidos pela minimização da função objetivo de mínimos quadrados penalizados,

$$\|y - X\beta\| + \lambda_1\beta^T S_1\beta + \lambda_2\beta^T S_2\beta$$

onde os parâmetros de suavização λ_1 e λ_2 controlam o grau de suavização de f_1 e f_2 , para que o ajuste final seja adequado aos dados. Por enquanto, assume-se que os parâmetros de suavização são conhecidos.

Define-se $S \equiv \lambda_1 S_1 + \lambda_2 S_2$, assim a função objetivo pode ser reescrita como:

$$\|y - X\beta\|^2 + \beta^T S\beta = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right\|^2$$

onde B é uma matriz raiz quadrado tal que $B^T B = S$. Apartir deste ponto o modelo pode ser ajustado por regressão linear simples.

3.4 Modelos Aditivos Generalizados

Modelos aditivos generalizados, são uma extensão do modelo aditivo, como os modelos lineares generalizados são do modelo linear, ver [Wood, 2006]. Neste enfoque o preditor linear passa a ser uma função suave conhecida para o valor esperado da variável resposta. A variável resposta agora pode seguir qualquer distribuição da família exponencial, ou ainda, definindo-se simplesmente um relacionamento médio conhecido permitindo o uso de estimação por quasi-verossimilhança. O modelo resultante é apresentado:

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots$$

onde

$$\mu_i \equiv E(Y_i) \quad \text{e} \quad Y_i \sim \text{família exponencial}$$

onde Y_i é a variável resposta, X_i^* é uma linha da matriz do modelo associada à parte estritamente paramétrica, θ é o vetor de parâmetros correspondentes, e as f_j são funções suaves das covariáveis x_k .

É claro que neste caso é necessário fazer as mesmas suposições da extensão do modelo linear para o linear generalizado. Também podemos pensar que como o modelo aditivo foi estimado por mínimos quadrados penalizados o modelo aditivo generalizado seria estimado por máxima verossimilhança. Na prática para ajustar o modelo é utilizado o método de mínimos quadrados re-ponderados penalizados iterativos (P-IRLS). Neste método um sistema de equações de estimação é iterado até a convergência dos valores dos parâmetros. O procedimento pode ser resumido da seguinte forma:

1. Dado o preditor linear corrente estimado, $\eta^{[k]}$, e as estimativas médias correspondentes ao vetor de resposta, $\mu_{[k]}$ calcula-se:

$$w_i \propto \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2} \quad \text{e} \quad z_i = g'(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + X_i\beta^{[k]}$$

onde $\text{var}(Y_i) = V(\mu^{[k]})\phi$ e X_i é uma linha da matriz X .

2. Minimizar

$$\left\| \sqrt{W}(z - X\beta) \right\|^2 + \lambda_1\beta^T S_1\beta + \lambda_2\beta^T S_2\beta$$

em relação a β para obter $\beta^{[k+1]}$. Então $\eta^{[k+1]} = X\beta^{[k+1]}$. W é uma matriz diagonal tal que, $W_{ii} = w_i$. O passo 2 pode ser dado equivalentemente por

Minimizar

$$\left\| \begin{bmatrix} \sqrt{W} & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} z \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right) \right\|^2$$

em relação a β para obter $\beta^{[k+1]}$, então $\eta^{[k+1]} = X\beta^{[k+1]}$. B é uma matriz raiz quadrada tal que $B^T B = \lambda_1 S_1 + \lambda_2 S_2$.

Agora só falta especificar como deve ser usado o escore da validação cruzada generalizada (GCV - *generalised cross validation*) para este

modelo. A forma mais natural e fácil é usar o GCV para o modelo linear resultante do P-IRLS iterativo. Existem outras formas de implementação e para mais detalhes ver [Wood, 2006, cap. 4].

Todos os procedimentos acima descritos estão implementados no pacote *mgcv* [Wood, 2008] do software **R** [Development Core Team, 2007], o qual foi utilizado para ajustar os modelos.

4 Resultados

Nesta seção será feita a apresentação dos principais resultados encontrados nas análises. A seção está dividida em duas partes, a primeira tratando da descrição da variável resposta, as estatísticas descritivas das covariáveis das armadilhas e a construção dos fatores climáticos utilizando a análise fatorial. Na segunda segunda parte é apresentada uma seleção dos resultados principais obtidos com a aplicação de modelos estatísticos relacionando a contagem de ovos com variáveis locais e ambientais.

4.1 Análise descritiva e construção dos fatores

O experimento de Recife/PE foi desenvolvido em seis bairros, porém um deles foi excluído da análise apresentada aqui, devido apresentar características distintas tais como ações de controle executadas, como inundação de ovitrampas. A análise reportada aqui refere-se então a cinco bairros e as abreviaturas de seus nomes listados na Tabela 1. Devido a problemas operacionais nas coletas tanto da variável resposta – contagem de ovos – como das covariáveis, principalmente climáticas, os bairros não apresentam o mesmo número de observações. A Tabela 1 resume a estrutura dos dados utilizados neste relatório.

Bairro	Obs.	Arma.	Início	Término	Sem.
BT: Brasília Teimosa	2580	80	04/01/2005	15/05/2007	124
CFP: Casa Forte	1800	100	04/01/2006	16/05/2006	72
DI: Dois Irmãos	1512	84	04/01/2005	16/05/2006	72
EM: Engenho do Meio	3225	100	04/01/2005	15/05/2007	124
MCP: Morro Conceição	1800	100	04/01/2005	16/05/2006	72

Tabela 1: Descrição do experimento por bairros.

A distribuição da variável resposta em cada um dos bairros pode ser vista, através dos histogramas e gráficos de densidade estimada da Figura 3, em duas escalas, logarítmica (esquerda) e original (direita). De forma geral

destacam-se nos gráficos a presença de contagens nulas e a forte assimetria causada por valores de contagem muito elevados, especialmente em CFP, DI e MCP.

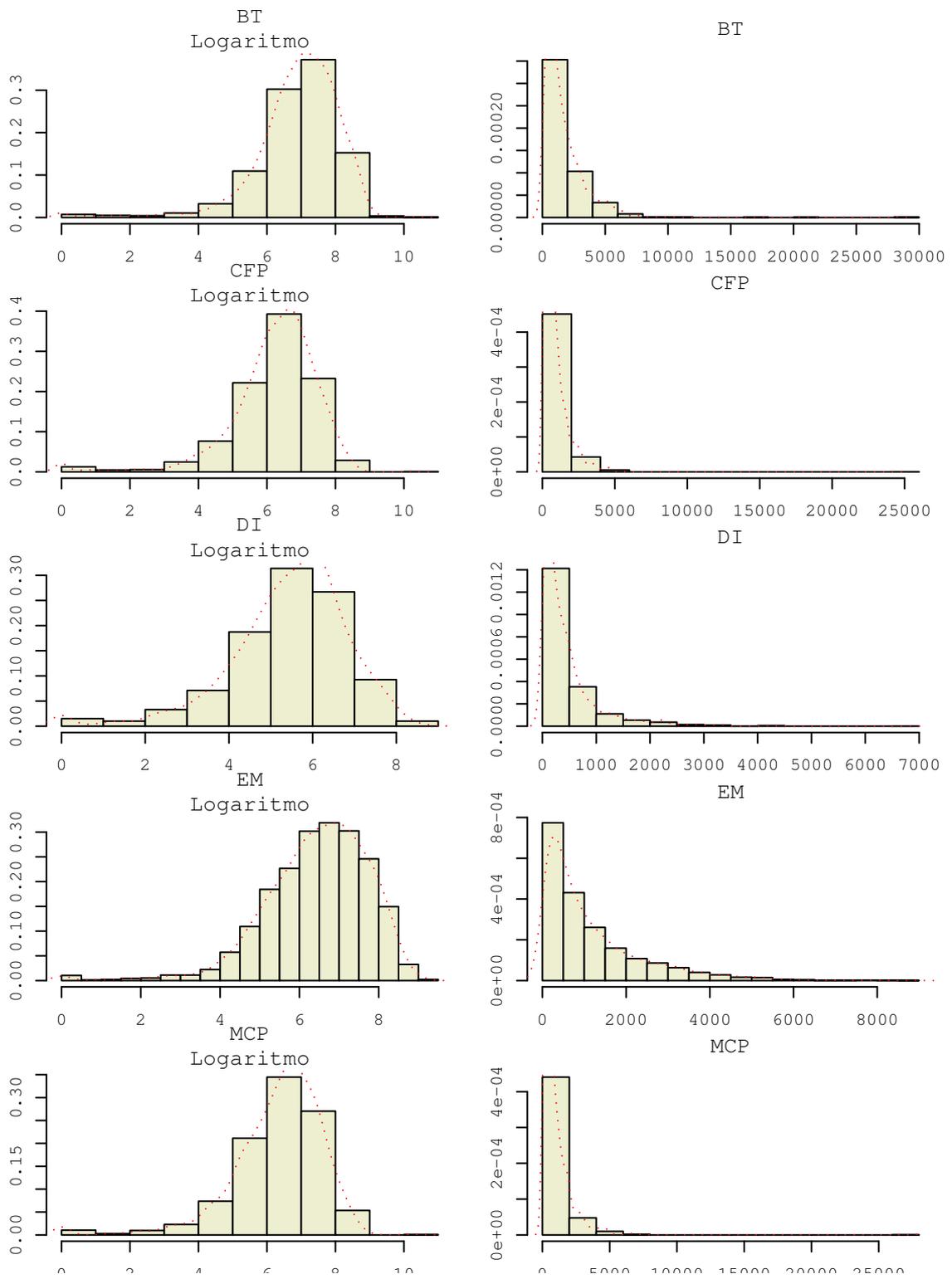


Figura 3: Histogramas e densidades estimadas em escala logarítmica (esquerda) e original (direita) do número de ovos nos bairros do estudo.

As condições ligadas às armadilhas são de papel fundamental pois são elas que orientam as políticas de prevenção à propagação do vetor através de campanhas educacionais promovidas afim de evitar criadouros do mosquito. Assim, a identificação das covariáveis realmente significativas é de fundamental importância para orientar as ações de tais campanhas. O conjunto de cinco tabelas correspondentes a cada um dos bairros apresentado a seguir, faz um resumo numérico comparando as contagens de ovos entre as duas categorias de cada uma das dez covariáveis consideradas neste trabalho. Vale lembrar que a descrição de cada covariável encontra-se na Seção 2.

A análise destes resultados permite identificar, para cada bairro de forma exploratória, os fatores que mais afetam as contagens de ovos, orientando a seleção e a escolha de modelos, antecipando e explicando possíveis resultados da modelagem. É ainda importante avaliar casos onde as diferenças são influenciadas pela presença de contagens muito elevadas.

Tabela 2: Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Brasília Teimosa.

Covariável	Cod	Min.	1 Qua.	Median	Mean	3 Qua.	Max.
Água Rede	0	0	584.2	1192	1708	2293	28420
Água Rede	1	3	469.8	826	1251	1752	5374
Canalizada	0	0	593.0	1193	1708	2277	28420
Canalizada	1	0	462.0	1028	1538	2120	7933
Abastecimento	0	0	495.2	1304	1571	2019	6619
Abastecimento	1	0	575.0	1163	1682	2260	28420
Quintal	0	0	586.0	1181	1684	2292	20760
Quintal	1	0	538.5	1141	1667	2086	28420
Grande c	0	0	543.0	1154	1784	2527	20760
Grande c	1	0	583.0	1171	1665	2219	28420
Grande s	0	0	563.0	1154	1674	2168	28420
Grande s	1	0	579.0	1187	1683	2332	20760
Pequeno c	0	0	557.0	1160	1659	2230	28420
Pequeno c	1	27	630.0	1188	1771	2378	20760
Pequeno s	0	0	575.0	1166	1677	2240	28420
Pequeno s	1	27	568.0	1191	1705	2364	6551
Fator risco	0	0	611.8	1248	1768	2340	28420
Fator risco	1	0	537.0	1062	1586	2168	20760
Tipo imóvel	0	0	585.2	1168	1675	2257	28420
Tipo imóvel	1	0	144.8	1168	1825	2292	17500

Tabela 3: Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Casa Forte / Parnamirin.

Covariável	Cod	Min.	1 Qua.	Median	Mean	3 Qua.	Max.
Água Rede	0	0	285.0	599.0	868.8	1136.0	24420
Água Rede	1	2	385.0	666.0	856.7	1030.0	4703
Canalizada	0	0	288.0	594.0	857.9	1120.0	24420
Canalizada	1	24	489.0	788.0	1028.0	1396.0	3636
Abastecimento	0	0	297.5	622.5	897.5	1136.0	24420
Abastecimento	1	0	291.0	602.5	849.3	1123.0	5378
Quintal	0	0	298.2	625.5	875.8	1136.0	24420
Quintal	1	29	217.8	425.0	713.8	914.5	4518
Grande c	0	0	413.8	784.5	1103.0	1602.0	5378
Grande c	1	0	277.0	580.5	820.6	1057.0	24420
Grande s	0	0	308.5	610.5	947.3	1203.0	24420
Grande s	1	0	290.0	611.0	842.3	1106.0	5378
Pequeno c	0	0	293.8	612.0	874.0	1133.0	24420
Pequeno c	1	12	291.8	549.0	696.6	1046.0	2066
Pequeno s	0	0	291.0	604.5	847.8	1110.0	24420
Pequeno s	1	12	423.2	935.0	1342.0	1738.0	5378
Fator risco	0	0	304.2	652.0	947.9	1175.0	24420
Fator risco	1	0	273.8	568.0	798.6	1076.0	4602
Tipo imóvel	0	0	323.0	663.0	925.7	1208.0	5378
Tipo imóvel	1	0	250.0	523.0	776.5	1004.0	24420

Tabela 4: Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Dois Irmãos.

Covariável	Cod	Min.	1 Qua.	Median	Mean	3 Qua.	Max.
Água Rede	0	0	117.0	283.0	482.7	566.2	6828
Água Rede	1	0	72.0	168.0	366.7	473.0	3373
Canalizada	0	0	108.0	263.0	464.2	551.2	4484
Canalizada	1	0	124.0	339.0	509.6	615.0	6828
Abastecimento	0	0	108.0	261.0	449.4	542.0	6828
Abastecimento	1	0	149.5	365.5	597.5	658.8	4310
Quintal	0	0	108.0	266.0	467.3	552.0	6828
Quintal	1	5	141.5	323.5	499.3	612.2	3078
Grande c	0	0	108.0	266.0	475.8	542.0	6828
Grande c	1	0	111.8	269.5	466.4	581.8	4310
Grande s	0	0	116.0	273.0	484.6	585.0	4484
Grande s	1	0	102.0	264.0	445.7	522.0	6828
Pequeno c	0	0	108.0	259.0	435.8	522.0	4310
Pequeno c	1	0	117.0	314.0	551.7	631.0	6828
Pequeno s	0	0	104.0	256.0	443.7	528.0	6828
Pequeno s	1	5	249.8	473.5	723.2	945.2	4156
Fator risco	0	0	135.0	289.0	476.2	587.5	4156
Fator risco	1	0	105.0	263.5	467.2	552.0	6828
Tipo imóvel	0	0	110.0	269.0	471.6	558.0	6828
Tipo imóvel	1	28	112.2	208.5	313.5	412.2	890

Tabela 5: Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Morro da Conceição.

Covariável	Cod	Min.	1 Qua.	Median	Mean	3 Qua.	Max.
Água Rede	0	0	290.0	678.0	1005.0	1316	27820
Água Rede	1	98	700.0	1324.0	1288.0	1571	2995
Canalizada	0	0	288.0	678.0	1016.0	1353	27820
Canalizada	1	50	422.8	741.5	852.9	1150	3028
Abastecimento	0	0	258.5	632.5	898.9	1225	6648
Abastecimento	1	0	327.0	752.0	1104.0	1483	27820
Quintal	0	0	301.0	689.0	1012.0	1340	27820
Quintal	1	0	241.8	653.5	978.9	1286	4881
Grande c	0	0	230.5	570.0	856.4	1162	5466
Grande c	1	0	307.8	727.5	1048.0	1410	27820
Grande s	0	0	295.0	666.0	1005.0	1315	27820
Grande s	1	0	289.5	729.0	1020.0	1408	5708
Pequeno c	0	0	300.0	729.0	1068.0	1462	27820
Pequeno c	1	0	284.8	643.5	925.7	1200	7407
Pequeno s	0	0	282.0	673.0	1010.0	1351	27820
Pequeno s	1	0	390.8	731.0	1000.0	1302	5708
Fator risco	0	0	309.0	733.0	1040.0	1426	7407
Fator risco	1	0	283.8	655.5	987.3	1288	27820
Tipo imóvel	0	0	291.0	681.0	1009.0	1324	27820
Tipo imóvel	1	157	443.0	926.0	998.0	1467	2752

Tabela 6: Estatísticas descritivas das contagens de ovos por categoria das covariáveis no Bairro Engenho do Meio.

Covariável	Cod	Min.	1 Qua.	Median	Mean	3 Qua.	Max.
Água Rede	0	0	303.0	744.0	1170.0	1614	8925
Água Rede	1	0	155.0	405.0	935.8	999	5229
Canalizada	0	0	300.8	734.5	1164.0	1598	8925
Canalizada	1	0	242.0	574.5	1112.0	1547	5229
Abastecimento	0	0	173.5	479.0	883.2	1378	5002
Abastecimento	1	0	303.0	735.0	1171.0	1600	8925
Quintal	0	0	292.0	699.0	1128.0	1534	8925
Quintal	1	0	415.5	1142.0	1556.0	2466	6727
Grande c	0	0	319.8	755.0	1204.0	1543	6379
Grande c	1	0	292.0	722.0	1154.0	1610	8925
Grande s	0	0	296.0	756.0	1182.0	1689	8162
Grande s	1	0	300.0	695.0	1142.0	1531	8925
Pequeno c	0	0	289.0	689.0	1124.0	1549	8925
Pequeno c	1	0	368.0	904.0	1307.0	1748	8124
Pequeno s	0	0	298.5	717.5	1153.0	1591	8925
Pequeno s	1	22	301.2	797.0	1221.0	1678	5817
Fator risco	0	0	366.2	810.5	1275.0	1755	7645
Fator risco	1	0	278.0	694.0	1130.0	1545	8925
Tipo imóvel	0	0	328.0	776.0	1210.0	1658	8925
Tipo imóvel	1	0	137.5	328.0	599.6	674	6727

Na modelagem considerada posteriormente é utilizada a inclusão no modelo das covariáveis referentes ao clima, através de fatores. Nas análises foram consideradas e avaliadas algumas estratégias para implementação da análise fatorial e a seguir descreve-se a escolhida.

Em um primeiro momento dividiu-se as covariáveis em dois fatores um denominado **Temperatura** que considera as covariáveis temperatura máxima e mínima, e um segundo fator denominada **Umidade** que considera as covariáveis umidade do ar e precipitação. Após o ajuste do modelo fatorial foram extraídas as cargas fatoriais e os fatores foram compostos conforme sua contribuição. As cargas fatoriais foram padronizadas para somarem 1, assim cada covariável tem uma contribuição para seu devido fator que pode ser representada em termos de percentuais. Tal representação é apresentada nas tabelas 7 e 8, para os fatores **Temperatura** e **Umidade**, respectivamente.

Em ambos casos considerou-se como variáveis medidas de 12 semanas que antecederiam a coleta de ovos. Na tabela 7 as variáveis correspondem às temperaturas mínimas (acima) e máximas (abaixo) de cada semana, enquanto que na tabela 8 considerou-se a precipitação (acima) e umidade (abaixo) média de cada uma das 12 semanas.

Cabe ressaltar que para o bairro Engenho do Meio o fator umidade não pôde ser composto, devido ao fato de que as observações em semanas sequenciais não apresentaram alta correlação. Sendo assim neste caso foram incluídas no modelo semana a semana sem a utilização dos fatores.

Os resultados permitiram agrupar nos fatores semanas consecutivas e, de forma geral, em grupos de três a cinco semanas, sugerindo que estas variáveis podem ser tomadas mensalmente. As cargas fatoriais variam um pouco dentro de cada fator, porém os resultados sugerem que seja investigada se tal variação necessita, de fato, ser tomada em consideração ou se uma simples medida mensal seria suficiente.

Tabela 7: Composição dos fatores de **Temperatura** em percentuais por bairro.

Semanas	BT	Fator	CFP	Fator	DI	Fator	MCP	fator	EM	Fator
Min 1	12.09	1	11.96	1	12.89	1	11.64	1	12.10	1
2	13.27	1	12.67	1	13.60	1	12.55	1	13.28	1
3	11.90	1	11.06	1	11.72	1	11.05	1	11.89	1
4	9.23	1	8.04	1	8.40	1	7.95	1	9.26	1
5	9.97	2	11.80	2	10.22	2	11.56	2	10.11	2
6	13.98	2	17.58	2	16.72	2	17.16	2	13.88	2
7	20.50	2	19.20	2	16.22	2	19.44	2	19.88	2
8	13.43	2	12.69	2	11.39	2	13.05	2	13.30	2
9	8.04	3	8.76	3	8.63	3	8.94	3	7.86	3
10	11.48	3	11.89	3	12.86	3	12.21	3	11.39	3
11	13.00	3	13.86	3	14.06	3	14.28	3	12.82	3
12	14.61	3	14.42	3	14.55	3	14.84	3	14.40	3
Max 1	15.02	1	14.06	1	15.68	1	14.28	1	14.94	1
2	15.18	1	13.67	1	14.63	1	13.86	1	15.24	1
3	12.87	1	11.50	1	12.45	1	11.61	1	12.95	1
4	10.43	1	9.60	1	10.63	1	9.75	1	10.33	1
5	12.32	2	7.45	1	11.07	2	7.30	1	12.41	2
6	16.00	2	13.99	2	12.26	2	14.18	2	16.20	2
7	13.79	2	13.91	2	12.02	2	13.74	2	14.21	2
8	6.27	3	10.83	2	10.11	2	10.87	2	6.26	3
9	8.98	3	9.89	3	9.83	3	9.47	3	9.12	3
10	11.53	3	12.07	3	11.79	3	11.85	3	11.66	3
11	13.29	3	14.43	3	14.00	3	14.15	3	13.53	3
12	12.79	3	14.67	3	14.28	3	14.26	3	12.97	3

Tabela 8: Composição dos fatores de **Umidade** em percentuais por bairro.

Semanas	BT	Fator	CFP	Fator	DI	fator	MCP	Fator
Precip 1	19.52	1	18.15	1	18.44	1	14.49	1
2	18.36	1	17.72	1	17.98	1	14.82	1
3	12.76	1	13.42	1	12.26	1	13.23	1
4	20.86	2	14.11	2	16.65	2	9.35	1
5	22.45	2	18.90	2	21.49	2	14.75	2
6	18.30	2	17.46	2	18.60	2	18.91	2
7	18.60	3	22.18	3	20.99	3	18.61	2
8	19.50	3	31.01	3	25.35	3	14.47	2
9	16.16	3	28.36	3	22.87	3	7.69	3
10	9.54	4	8.91	4	8.28	4	11.34	3
11	13.92	4	14.31	4	14.01	4	13.47	3
12	16.08	4	16.77	4	16.08	4	13.72	3
Umid 1	19.77	1	18.14	1	18.48	1	13.10	1
2	17.59	1	18.51	1	18.87	1	13.47	1
3	12.00	1	14.06	1	13.97	1	12.14	1
4	20.22	2	11.03	2	12.81	2	9.40	1
5	18.16	2	15.25	2	16.19	2	9.76	2
6	14.14	3	13.79	2	14.27	2	11.88	2
7	16.88	3	9.45	2	14.15	3	11.62	2
8	14.74	3	18.45	3	16.64	3	7.27	3
9	10.58	4	9.69	4	10.08	4	9.76	3
10	14.83	4	14.40	4	14.59	4	11.73	3
11	17.67	4	17.91	4	18.39	4	12.63	3
12	17.38	4	18.00	4	18.56	4	12.40	3

4.2 Aplicação do modelo estatístico

Para determinar quais são os fatores de risco e proteção para a ocorrência de ovos do mosquito *Aedes aegypti* foram utilizados modelos lineares generalizados e modelos aditivos generalizados. A distribuição de probabilidade assumida para a variável resposta foi a binomial negativa com função de ligação logarítmo. Em comparações não reportadas aqui a binomial negativa apresentou resultados claramente superiores aos obtidos com a distribuição de Poisson. Nas análises mostradas a seguir, foram ajustados, para cada um dos bairros separadamente, seis modelos. Partiu-se do modelo com efeitos mais simples até chegar aos efeitos mais complexos como efeitos espaço-temporais. A denominação e a composição de cada um dos modelos é dada a seguir.

1. **Completo Covariáveis** - É o modelo com todas as 10 covariáveis referentes às características das armadilhas.
2. **Selecionado Covariáveis** - Com base no modelo Completo Covariáveis foram selecionadas as covariáveis significativas, através de um processo do tipo *StepWise*, o valor de corte foi p-valor 0.10.
3. **Completo Climáticas** - É o modelo com todos os fatores referentes às covariáveis climáticas.
4. **Selecionado Climáticas** - Com base no modelo Completo Climáticas foram selecionados os fatores significativos, através de um processo do tipo *StepWise*, o valor de corte foi p-valor 0.10.
5. **Combinado** - É o modelo resultante da combinação entre as covariáveis das armadilhas e os fatores climáticos, conforme os modelos 2 e 4.
6. **Espaço - temporal** - Com base no modelo Combinado, foram incluídos os efeitos espaciais através das coordenadas geográficas de cada armadilha, e o efeito temporal através de uma sequência suave das datas de coleta.

Devido ao grande número de covariáveis e também o número de modelos ajustados, optou-se por apresentar nesta seção os resultados do modelo Combinado e analisar os resultados conjuntamente. O ajuste de cada um dos modelos citados é apresentado em anexo. A comparação dos modelos para escolha do melhor ajuste será realizada pelo Critério de *Akaike*[Akaike, 1974]. Após a escolha do melhor ajuste será apresentado o modelo escolhido em detalhes e uma análise gráfica dos resíduos para a verificação dos pressupostos.

A Tabela 9 apresenta os resultados do ajuste do modelo Combinado para os cinco bairros em análise, lembrando que para o bairro Engenho do Meio o fator UMIDADE não está sendo considerado.

Tabela 9: Estimativa e p -valor do modelo combinado para todos os bairros.

Bairro	BT		CFP		DI		MCP		EM	
	Estimativa	p -valor								
Intercepto	5.947	$< 2e^{-16}$	1.235	0.073	-1.496	0.045	1.787	0.010	6.099	$4.1e^{-08}$
Água rede	-0.345	0.000	-	NS	-0.280	0.004	-	NS	-	NS
Canalizada	-0.128	0.041	-	NS	NS	-	-	NS	-	NS
Abastecimento	-	NS	-	NS	0.432	$9.9e^{-07}$	0.240	$1.1e^{-05}$	-	NS
Quintal	-	NS	-0.253	0.039	-	NS	-	NS	0.402	$1.2e^{-08}$
Grande c	-	NS								
Grande s	-	NS								
Pequeno c	-	NS	-	NS	0.166	0.011	-0.206	0.000	0.096	0.042
Pequeno s	-	NS	0.386	0.004	0.572	$3.1e^{-08}$	-	NS	-	NS
Fator risco	-0.170	0.000	-0.093	0.088	-	NS	-	NS	-	NS
Tipo imóvel	-	NS	-0.132	0.015	-	NS	-	NS	-0.762	$< 2e^{-16}$
Umidade 1	-	NS	0.021	0.001	-	NS	0.047	$1.4e^{-13}$	-	-
Umidade 2	0.022	$4.9e^{-05}$	0.017	0.004	0.060	$< 2e^{-16}$	0.029	$2.8e^{-07}$	-	-
Umidade 3	0.023	$2.6e^{-05}$	-	NS	-	NS	-	NS	-	-
Umidade 4	-	NS	-	NS	-	NS	-	NS	-	-
Temperatura 1	-	NS	-	NS	-	NS	-	NS	-0.206	$4.1e^{-08}$
Temperatura 2	-	NS								
Temperatura 3	-	NS	0.148	$1.6e^{-08}$	0.197	$8.5e^{-13}$	0.085	0.001	0.083	$1.3e^{-06}$

A covariável água na rede, apresentou-se significativa para os bairros Brasília Teimosa e Dois Irmãos, a presença de água ligada à rede geral é um fator de risco. Fato este curioso, pois era de se esperar que à água ligada a rede geral indica-se melhores condições sociais, e conseqüente queda na variável resposta devido ao maior cuidado e informação da população sobre os criadouros do mosquito. Uma possível explicação para tal resultado, seria que para as populações não servidas pela rede geral de água, tal benefício escasso é tratado com maior cuidado, e pouco armazenado evitando o desperdício e conseqüentemente criadouros do mosquito.

A covariável água canalizada apresentou-se significativa apenas para o bairro Brasília Teimosa, e deve ser considerada como um fator de risco, as mesmas explicações do caso anterior servem para esta covariável.

Com relação à frequência de abastecimento de água, verifica-se que para os bairros Dois Irmãos e Morro da Conceição, a falta de frequência no abastecimento de água é considerado um fator de risco, este fato é esperado, já que, devido a falta de abastecimento contínuo de água a população é obrigada a ter reservatórios no domicílio, aumentando assim a quantidade de água ofertada para o mosquito.

A presença de quintal ou jardim, mostrou-se significativa para os bairros Casa Forte/Parnamirin e Engenho do Meio, porém com efeitos contrários, no Casa Forte/Parnamirin a presença de quintal ou jardim é um fator de risco, já para o bairro Engenho do Meio é considerada um fator protetor. O fato de hora a covariável apresentar-se como risco e hora proteção deve estar ligada a condições de infraestrutura dos bairros que não estão sendo levados em consideração neste estudo.

A presença de recipientes pequenos com tampa, mostrou-se significativa para os bairros Dois Irmãos e Morro da Conceição, porém com efeitos diferentes, para o bairro Dois Irmãos é considerado fator de risco, já para o bairro Morro da Conceição mostra-se como fator protetor, o que era de se esperar, já que, a presença de jarros de barro ou baldes mesmo com tampa pode atrair uma quantidade maior de mosquitos. O fato da covariável apresentar-se com sinais trocados em bairros diferentes, pode ser devido a algum fator de confundimento tal como altitude, ou mudanças bruscas nas condições

sociais dentro de um mesmo bairro, fazendo com que a configuração espacial aponte tal covariável como significativa, quando na verdade ela está apenas expressando um retrato de uma configuração espacial ainda não levada em consideração.

A presença de recipientes pequenos sem tampa, mostrou-se significativa para os bairros Casa Forte/Parnamirin e Dois Irmãos, para os dois foi considerado um fator de proteção, este resultado deve ser encarado com cuidados porque o número de observações onde as armadilhas não estão perto de tais recipientes é muito pequeno, por exemplo para o bairro Casa Forte / Parnamirin tem-se na categoria (0) que é onde em sua proximidade encontram-se, jarros de barro ou baldes sem tampa, um total de 1728 observações, enquanto que para a categoria (1), que não apresentam tais características, apenas 72 observações. Para o bairro Dois Irmãos a categoria (0) contou com 1368 observações e a categoria (1) contou com apenas 144.

A covariável tipo de imóvel apresentou-se significativa para os bairros, Brasília Teimosa e Casa Forte/Parnamirin, nos dois casos os resultados mostram que o local preferido do mosquito para a colocação dos ovos são, casas, apartamentos e residências em geral, ao invés de, quartéis, escolas e outros.

Com relação aos fatores climáticos é possível observar que o fator Umidade 2 e temperatura 3, são os que apresentam significância em quatro dos cinco bairros em análise, o que permite inferir, que um período seco, consequentemente de altas temperaturas, seguido de um período de aproximadamente um mês para a reprodução do vetor, em seguida um período de alta umidade e consequente queda na temperatura, é a condição ideal para a ocorrência de ovos do mosquito *Aedes aegypti*.

O modelo **Combinado** apresentado acima, permitiu identificar quais são as covariáveis que se relacionam significativamente com a resposta. Em alguns casos, os resultados foram inconclusivos ou incoerentes com o que se esperava, mostrando que o ajuste de tal modelo não é suficiente para se tirar conclusões coerentes, já que, se está ignorando a condição espaço temporal em que o experimento foi realizado. Neste momento mostra-se de interesse fazer uma comparação dentre os modelos ajustados, para ver qual seria o modelo que melhor está descrevendo a relação entre a resposta e as covariáveis e as-

sim definir quais são as relações realmente importantes. A tabela 10 resume esta situação.

Tabela 10: Comparação entre os modelos ajustados através do Critério de *Akaike*.

Modelos	Bairros				
	BT	CFP	DI	MCP	EM
Completo Covariáveis	37125	23398	19889	22838	43255
Selecionado Covariáveis	37114	23414	19883	22820	43244
Completo Climática	37032	23220	19730	22579	--
Selecionado Climáticas	37044	23215	19722	22579	--
Combinado	36996	23191	19626	22526	42691
Espaço-temporal	36772	23066	19379	22431	42540

A superioridade do modelo com efeitos espaço-temporal é evidente para todos os bairros. Nota-se que a queda do AIC, não é muito grande de um modelo para o seguinte, o que evidencia a dificuldade da modelagem estatística, na busca de um bom modelo para descrever o comportamento da variável resposta. Também pode-se dizer que apenas as covariáveis climáticas e das armadilhas não são suficientes para explicar o comportamento da resposta, efeitos espaciais e temporais quando introduzidos no modelo melhoraram o ajuste significativamente, mostrando que ainda deve haver muitas covariáveis importantes que não estão sendo levadas em consideração neste trabalho, pode-se citar como exemplo, condições sócio econômicas e culturais, que devem ser refletidas através da significância do efeito espacial. Outro fator são tendências ou sazonalidades, muitas vezes decorrentes das estações do ano, mesmo o modelo tendo covariáveis climáticas, ainda mostrou-se altamente significativa, novamente refletindo a influência de outras covariáveis não levadas em consideração. Dado estes resultados o modelo dentre os ajustados que melhor explica a variável resposta (número de ovos de *Aedes aegypti*), é o modelo que contém os efeitos espaço-temporais, sendo assim, para completar a análise, deve-se fazer uma descrição detalhada deste modelo para cada um dos bairros, e depois proceder uma análise de resíduos, para realmente poder-se concluir sobre o resultado do ajuste. Para todas as estimativas que apresenta-se um intervalo de confiança este será de 95%.

4.2.1 Brasília Teimosa

O resultado do ajuste do Modelo Aditivo Generalizado, para o bairro Brasília Teimosa é apresentado na tabela 11, lembrando que a distribuição de probabilidade assumida para a variável resposta, é a binomial negativa, a função de ligação é o logarítmo, e a inclusão dos efeitos espaciais foi feita utilizando a *Thin Plate Spline*, para o efeito temporal foi usada a *Cubic Regression Spline*, porque o algoritmo não convergiu usando a *Thin Plate Spline*.

Covariáveis	Estimativas	Erro Padrão	Valor t	<i>p</i> -valor
Intercepto	5.809	0.146	39.52	$< 2e - 16$
UMIDADE.2	0.019	0.005	3.875	0.000
UMIDADE.3	0.027	0.005	5.431	$6.2e - 08$
Água rede	-0.254	0.107	-2.373	0.017
Fator risco	-0.153	0.052	-2.940	0.003
Canalizada	-0.286	0.074	-3.853	0.000
Suavização	<i>Edf</i>	Rank	F	<i>p</i> -valor
Coordenadas	26.56	29	9.924	$< 2e - 16$
Tempo	5.68	9	2.190	0.020

Tabela 11: Ajuste do Modelo Aditivo Generalizado para o bairro Brasília Teimosa.

Para as covariáveis climáticas os fatores de umidade 2 e 3 foram significativos, as defasagens importantes para este fator é de 4 a 9 semanas, mostrando que o mosquito tem preferência por alta umidade.

O fato da localização da armadilha estar ligada a rede geral de água é indicado como um fator de risco, ao comparar as duas categorias desta covariável observa-se que a ocorrência de ovos é em média 28% (4.53% 62.89%) maior. A presença de plantas em vasos e outras características consideradas de risco, se confirmou como tal, pode-se dizer que a presença de tais fatores aumenta a ocorrência em 16% (5.2% 29%), a condição de possuir água canalizada em pelo menos um cômodo, também considerada fator de risco, 33% (15% 53%), corrobora o resultado encontrado para a água ligada a rede geral. Esses resultados também podem ser verificados analisando as

estatísticas descritivas da tabela 2.

A estimativa da função suave das coordenadas geográficas, pode ser apresentada através de um mapa da densidade de ocorrência de ovos do mosquito, para a construção deste mapa controla-se as outras covariáveis do modelo, fazendo variar apenas as coordenadas, o resultado é um mapa onde podem ser visualizadas as áreas de maior ocorrência da resposta controlando-se pelas outras covariáveis do modelo. A figura 4 apresenta tal superfície, a construção foi realizada mantendo-se constante no nível médio as covariáveis climáticas, e as covariáveis categóricas foram mantidas na categoria de referência, o efeito temporal foi mantido em uma semana além da amostra.

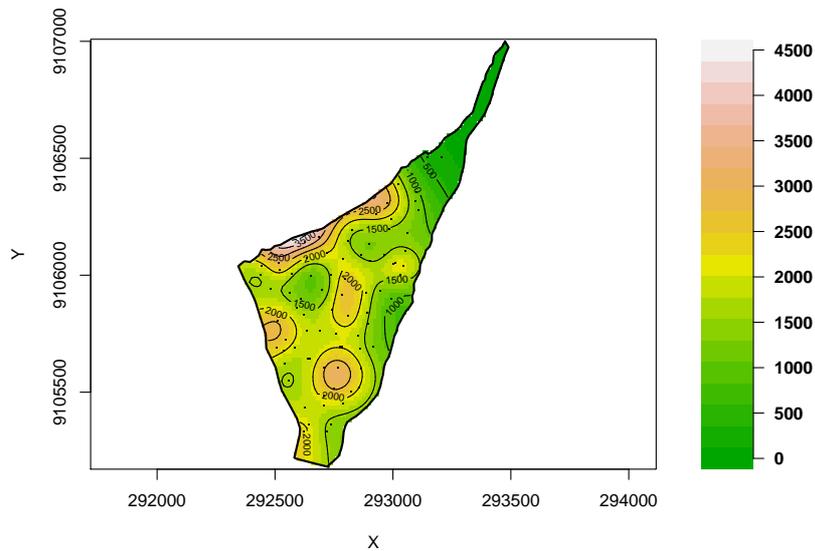


Figura 4: Superfície estimada da ocorrência de ovos de *Aedes aegypti* - Brasília Teimosa.

A estimativa do efeito temporal, é uma função suave das datas de observação, ele pode ser visto na figura 5, pode-se notar um pico começando no final do mês maio de 2006, tendo seu ponto máximo entre os meses de agosto e setembro, não é possível identificar sazonalidade na série, a tendência a partir de abril de 2007 mostra-se crescente, o que foi observado também para o ano em 2006, porém em 2005 o comportamento foi decrescente.

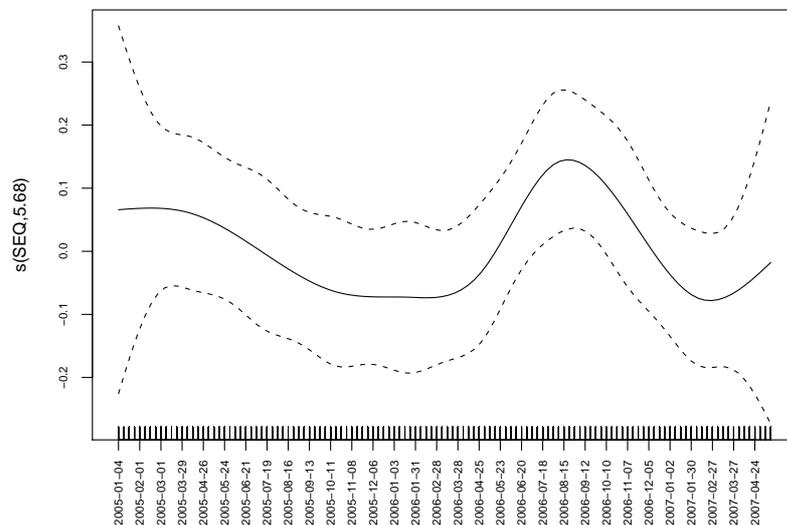


Figura 5: Ajuste do efeito temporal - Brasília Teimosa.

Da mesma forma, que os modelos lineares generalizados, tem uma série de pressupostos, o modelo aditivo também tem, sendo assim, a análise de pressupostos foi realizada através de técnicas gráficas baseadas nos resíduos do modelo, em geral espera-se que se o modelo é adequado os resíduos tenham distribuição Normal de média 0, sejam independentes e homocedásticos (variância constante). O conjunto de 4 gráficos apresentado na figura 6, resume esta situação.

Os gráficos mostram que nenhum dos pressupostos está sendo violado. Pelo gráfico Normal de probabilidade e o histograma dos resíduos, pode-se ver uma calda levemente mais pesada do que a esperada pela Normal, este fato é devido o grande número de observações iguais a 0, indicando que um número maior de zeros do que o esperado por uma distribuição Binomial Negativa foi observado, porém este fato não alterou as estimativas do modelo. Mas é um indicativo que a atribuição para a resposta de uma distribuição Binomial Negativa inflacionada de zeros, pode ser mais adequada do que a considerada neste trabalho.

Com respeito a valores discrepantes ou aberrantes que poderiam ser even-

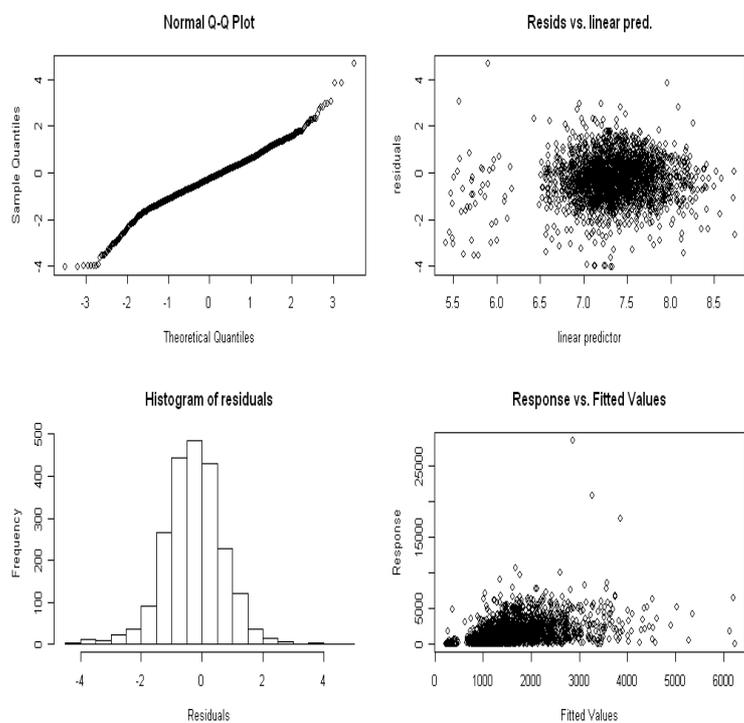


Figura 6: Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Brasília Teimosa.

tuais pontos influentes ou de alavancagem, não se encontrou nenhum. Pode-se então concluir que o modelo proposto é adequado.

4.2.2 Casa Forte / Parnamirin

O resultado do Modelo Aditivo Generalizado, para o bairro Casa Forte / Parnamirin é apresentado na tabela 12.

Covariáveis	Estimativas	Erro Padrão	Valor t	<i>p</i> -valor
Intercepto	1.232	0.634	1.942	0.052
TEMP.3	0.147	0.024	6.126	$1.1e - 09$
UMIDADE.1	0.018	0.006	3.020	0.002
UMIDADE.2	0.018	0.005	3.209	0.001
Pequeno s	0.435	0.138	3.138	0.001
Quintal	-0.388	0.127	-3.052	0.002
Suavização	<i>Edf</i>	Rank	F	<i>p</i> -valor
Coordenadas	24.30	29	5.089	$< 2e - 16$
Tempo	7.43	9	4.074	$3.51e - 05$

Tabela 12: Ajuste do Modelo Aditivo Generalizado para o bairro Casa Forte / Parnamirin.

Os fatores climáticos foram altamente significativos, o fator temperatura 3, mostra a preferência do vetor por altas temperaturas e indica que as semanas mais importantes são de 9 a 12 antes da observação. Para os fatores referentes a umidade 1 e 2, mostra a preferência por alta umidade as semanas mais importantes são de 1 a 6. Estes resultados indicam que um período de altas temperaturas e conseqüentemente seco, um período de aproximadamente 1 mês para a reprodução do vetor, e em seguida um período de alta umidade é a condição ideal para a ocorrência de ovos do mosquito.

A covariável presença de recipientes pequenos sem tampas mostra-se como um fator protetor, resultado não esperado, já que, recipientes como jarros de barro ou baldes são considerados possíveis criadouros do mosquito. Para explicar este resultado, primeiro, deve-se considerar a grande diferença no número de observações entre as categorias da covariável, para a categoria (0) que considera as localizações das armadilhas onde encontram-se jarros de barro ou baldes sem tampas, tem-se 1728 observações enquanto que para a categoria (1) tem-se apenas 72 observações. Segundo, este resultado pode ser um indicativo de que exista algum tipo de competição entre estes tipos

de recipientes e as armadilhas. Calculando a ocorrência proporcional tem-se que em média a categoria 1 é 54.4% (17.8% - 102%) maior que a categoria 0.

A presença de quintal ou jardim, mostra-se como um fator de risco, a ocorrência é 47.4% (14.9% - 89%) vezes maior que a categoria onde as localizações não apresentam quintal ou jardim. Os resultados para as covariáveis podem também ser confirmados pelas estatísticas descritivas apresentadas na tabela 3.

O ajuste da função suave das coordenadas geográficas, é apresentado através da superfície estimada da ocorrência de ovos, mantendo-se as outras covariáveis no modelo constantes. Para a construção do mapa de densidade, os fatores climáticos foram mantidos constantes na média, as covariáveis da armadilha na categoria de referência e a função suave do tempo em uma semana após a última observação, o resultado pode ser visto na figura 7.

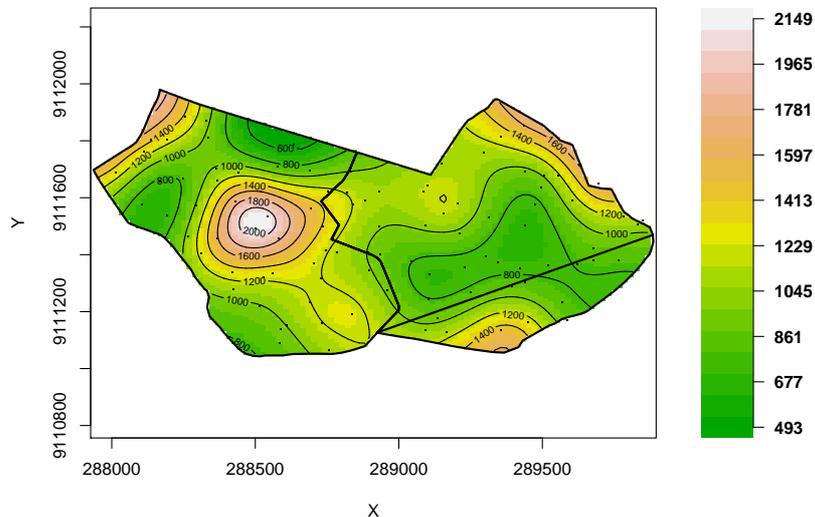


Figura 7: Superfície estimada da ocorrência de ovos de *Aedes aegypti* - Casa Forte / Parnamirin.

A estimativa do efeito temporal, é uma função suave das datas de observação, o ajuste pode ser visto na figura 8. Pode-se notar um padrão

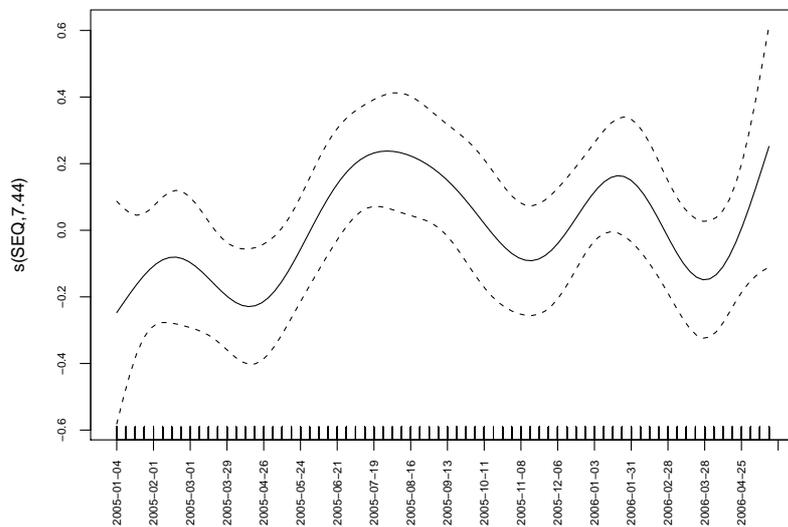


Figura 8: Ajuste do efeito temporal - Casa Forte / Parnamirin.

na série, sendo que um período de alta ocorrência é sempre seguido de um período de baixa ocorrência, dois grandes picos podem ser vistos entre os meses de julho e agosto, e em janeiro, a série em maio de 2005 teve um crescimento bastante acelerado, em maio de 2006 nota-se novamente a tendência de crescimento.

A análise de resíduos foi realizada através de técnicas gráficas, nota-se no histograma e no gráfico Normal de probabilidade, uma calda mais pesada do que a Normal, novamente devido ao grande número de zeros. Verifica-se também a presença de um ponto aberrante, porém devido ao número razoável de observações ele não tem grande influência sobre as estimativas dos β' s. Sendo assim conclui-se que o modelo é adequado para descrever a variável resposta.

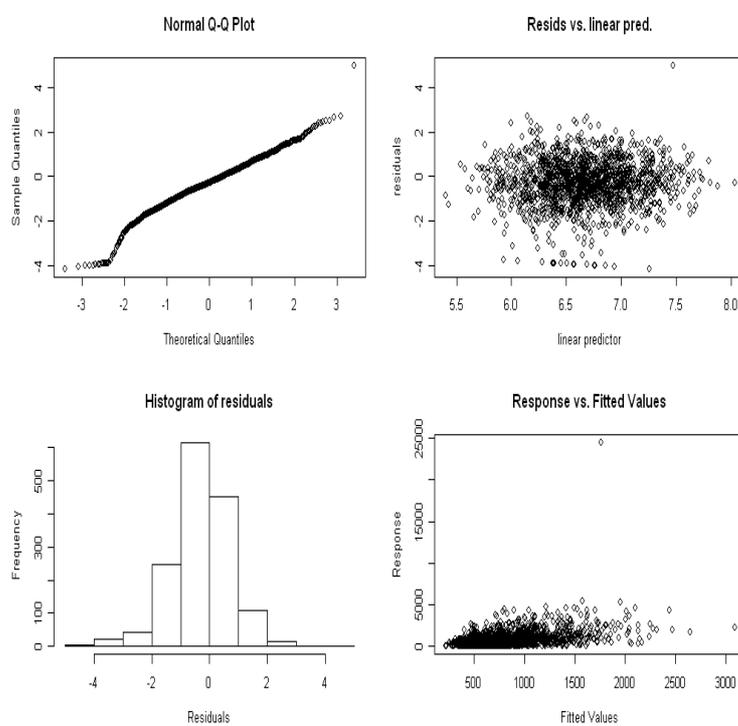


Figura 9: Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Casa Forte / Parnamirin.

4.2.3 Dois Irmãos

O resultado do ajuste do Modelo Aditivo Generalizado para o bairro Dois Irmãos é apresentado na tabela 13.

Covariáveis	Estimativas	Erro Padrão	Valor t	<i>p</i> -valor
Intercepto	-1.926	0.662	-2.906	0.003
TEMP.3	0.213	0.024	8.830	$< 2e - 16$
UMIDADE.2	0.058	0.004	13.082	$< 2e - 16$
Pequeno s	0.407	0.103	3.933	$8.82e - 05$
Abastecimento	0.264	0.109	2.407	0.016
Suavização	<i>Edf</i>	Rank	F	<i>p</i> -valor
Coordenadas	23.48	29	10.30	$< 2e - 16$
Tempo	7.77	9	3.74	0.000

Tabela 13: Ajuste do Modelo Aditivo Generalizado para o bairro Dois Irmãos.

Para os fatores climáticos verifica-se a significância da temperatura indicando uma defasagem de 9 a 12 semanas tem grande impacto na variável resposta. Para o fator umidade a defasagem significativa ocorreu de 4 a 6 semanas, indicando que um período de altas temperaturas, um período de aproximadamente um mês para a reprodução, seguido de um período de alta umidade perfaz a condição ideal para o aumento da ocorrência de ovos de *Aedes aegypti*. Ao avaliar-se as covariáveis das armadilhas, mostram-se significativas a presença de recipientes pequenos sem tampa, como fator protetor, já que, as armadilhas onde não se encontram tais recipientes têm em média 50.3% (22.7% 84.2%) mais ovos quando comparado as localizações que tem em suas proximidades tais objetos. Este fato pode indicar que um efeito de competição existe entre tais recipientes e as ovitrampas. A frequência do abastecimento de água mostrou-se significativa, é considerado um fator protetor, as localizações das armadilhas que não tem abastecimento de água apresentam em média 30.2% (5.02% 61.5%) mais ovos quando comparada com as localizações onde o abastecimento é diário. Resultado esperado, pois localizações que ainda não são servidas de abastecimento de água diário, são em geral zonas bastante carentes, onde as condições de saneamento básico ainda não foram implementadas pelo poder público, são também zonas de

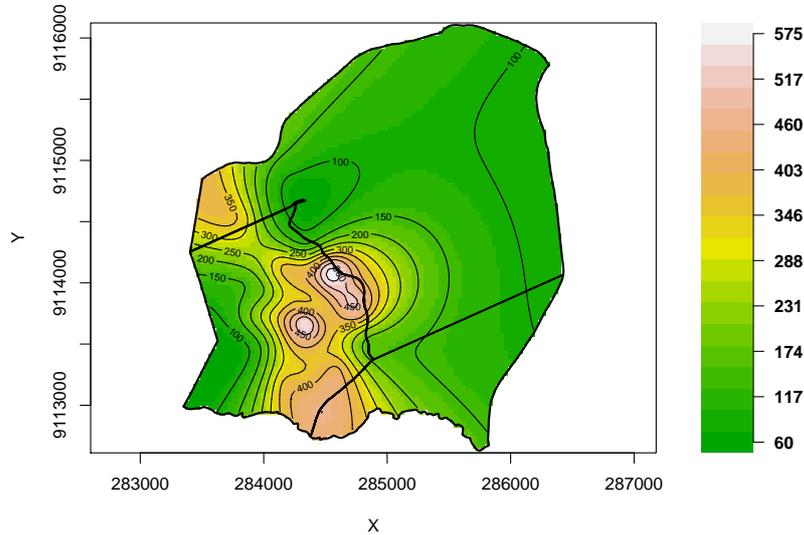


Figura 10: Superfície estimada da ocorrência de ovos de *Aedes aegypti* - Dois Irmãos.

baixa renda e baixo nível educacional, também o fato da água não ser disponibilizada diariamente aumenta os reservatórios dando maior possibilidade do mosquito se desenvolver.

O ajuste da função suave das coordenadas geográficas, é apresentado através da superfície estimada da ocorrência de ovos. Para a construção do mapa de densidade, os fatores climáticos foram mantidos constantes na média, as covariáveis da armadilha na categoria de referência e a função suave do tempo em uma semana após a última observação, o resultado pode ser visto na figura 10.

A estimativa do efeito temporal, também altamente significativa, é uma função suave das datas de observação, pode-se verificar um grande pico em julho de 2005, seguido por uma queda continua até fevereiro de 2006, quando a série volta a crescer, neste período em 2005 também verificou-se um aumento no nível geral da série. O ajuste do efeito temporal é apresentado na figura 11.

Para verificar a qualidade do ajuste foi realizada uma análise gráfica dos

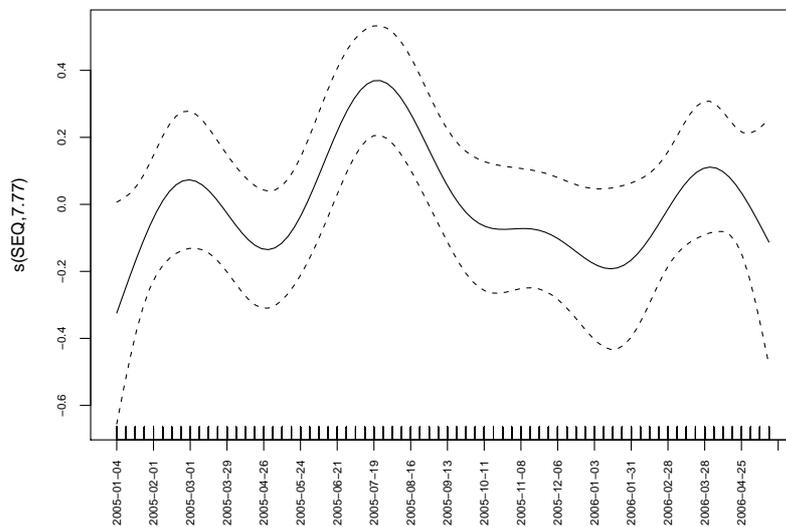


Figura 11: Ajuste do efeito temporal - Dois Irmãos.

resíduos, para identificar possíveis quebras dos pressupostos. O conjunto de quatro gráficos apresentado na figura 12 resume esta situação, de acordo com os resultados dos gráficos nenhum dos pressupostos está sendo violado, indicando que o modelo está bem ajustado.

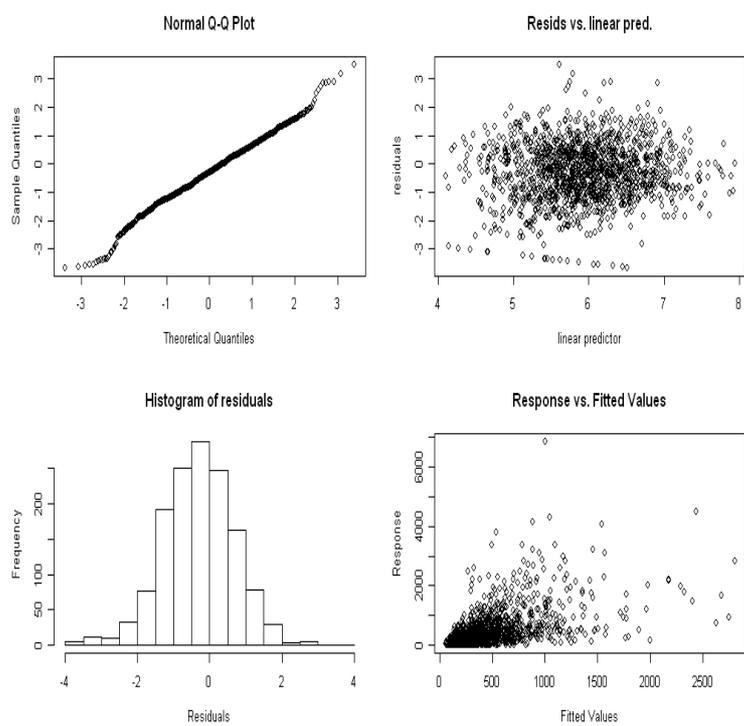


Figura 12: Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Dois Irmãos.

4.2.4 Morro da Conceição

O resultado do ajuste do Modelo Aditivo Generalizado para o bairro Morro da Conceição é apresentado na tabela 14.

Covariáveis	Estimativas	Erro Padrão	Valor t	<i>p</i> -valor
Intercepto	1.616	0.665	2.427	0.015
UMIDADE.1	0.044	0.006	7.375	$2.80e - 13$
UMIDADE.2	0.032	0.005	5.943	$3.52e - 09$
TEMP.3	0.092	0.026	3.570	0.000
Suavização	<i>Edf</i>	Rank	F	<i>p</i> -valor
Coordenadas	24.15	29	3.94	$1.64e - 11$
Tempo	8.875	9	3.53	0.000

Tabela 14: Ajuste do Modelo Aditivo Generalizado para o bairro Morro Conceição.

Para o bairro Morro da Conceição nenhuma das covariáveis da armadilha mostrou-se significativa, quando incluídos os efeitos espaciais e temporais. Para os fatores climáticos verifica-se que a defasagem para o fator temperatura é de 9 a 12 semanas, para o fator umidade a defasagem é de 1 a 6 semanas antes da observação. Novamente um período de altas temperaturas seguido de um período de aproximadamente 4 semanas para a reprodução do vetor, e um período de alta umidade perfazem a condição ótima para a ocorrência de ovos do mosquito.

Para este bairro o efeito espacial, mostrou-se altamente significativo, o ajuste pode ser visto na figura 13. Para a construção da superfície, os fatores climáticos foram mantidos constantes na média, o tempo é constante em uma semana após a última observação.

O efeito temporal foi expresso por uma função suave das datas de observação, verifica-se uma alta volatilidade na série, os meses de maior ocorrência de ovos são nos meses de fevereiro, junho, outubro. Verifica-se que os picos ocorrem de quatro em quatro meses.

Para verificar o ajuste do modelo procedeu-se uma análise gráfica dos resíduos, pode-se verificar no histograma e no gráfico Normal de Probabilidade uma calda levemente mais pesada, principalmente à esquerda do

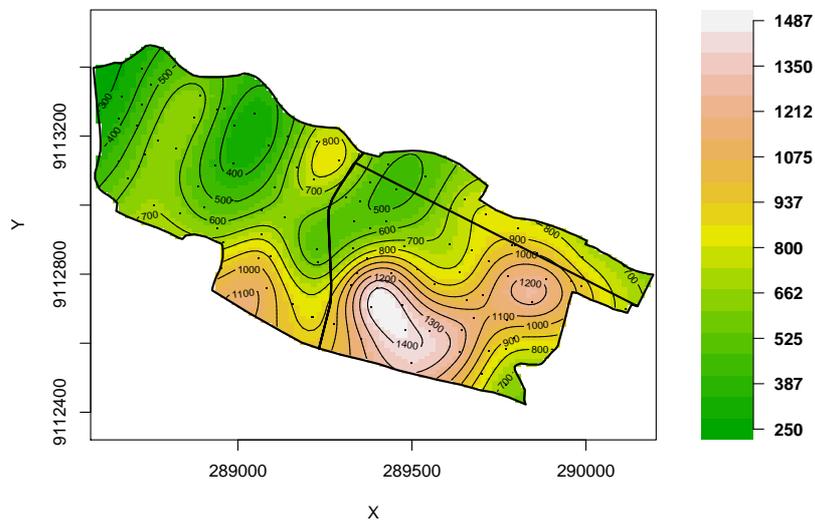


Figura 13: Superfície estimada da ocorrência de ovos de *Aedes aegypti* - Morro da Conceição.

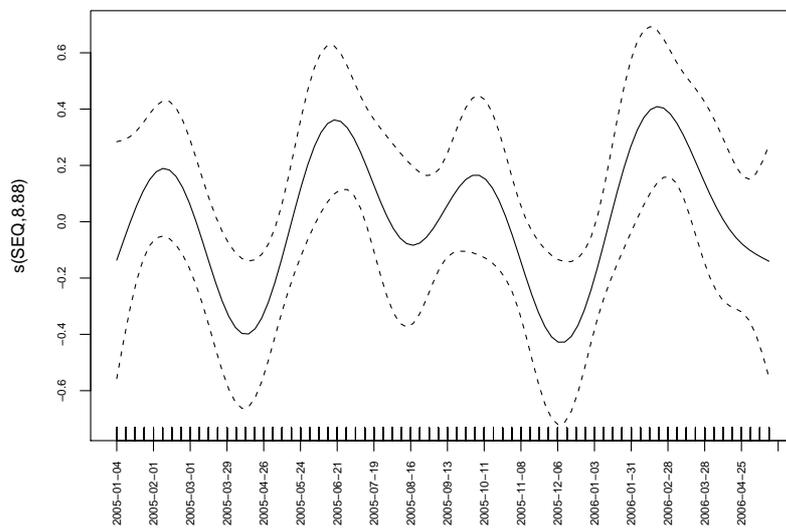


Figura 14: Ajuste do efeito temporal - Morro da Conceição.

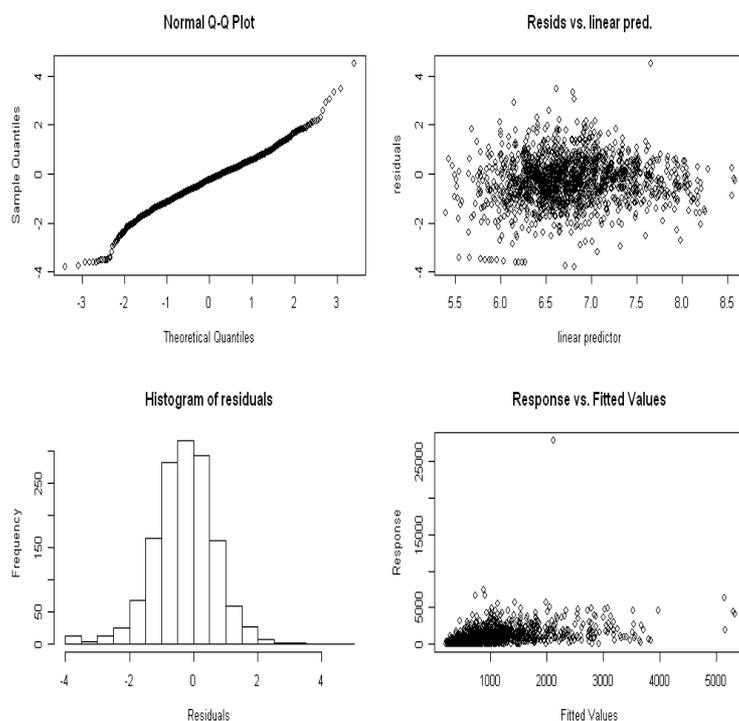


Figura 15: Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Morro da Conceição.

que a esperada por uma distribuição Normal padrão, novamente este fato é decorrência do grande número de zeros observados, mesmo assim a grande maioria dos pontos encontram-se sobre a reta e no centro do gráfico, indicando que a suposição de Normalidade não está sendo violada. Com relação a pontos que possam estar influenciando demasiadamente a estimação dos β 's que se caracterizariam como influentes, alavancagem ou aberrantes, não se encontrou nenhum, isto decorrente da quantidade considerável de observações utilizadas para ajustar o modelo. O conjunto de quatro gráficos apresentados na figura 15 resume todas estas informações sobre o ajuste do modelo, sendo assim pode-se concluir que o modelo está adequado.

4.2.5 Engenho do Meio

O resultado do ajuste do Modelo Aditivo Generalizado para o bairro Engenho do Meio é apresentado na tabela 15. Lembrando que para este bairro o fator umidade não pode ser composto, portanto o ajuste leva em consideração as covariáveis umidade e precipitação como a média da respectiva semana, com uma defasagem de até 12 semanas.

Covariáveis	Estimativas	Erro Padrão	Valor t	<i>p</i> -valor
Intercepto	5.601	1.076	5.201	$2.13e - 07$
TEMP.1	-0.1936	0.041	-4.661	$3.31e - 06$
TEMP.3	0.070	0.035	1.988	0.046
UMID.MED.1	0.019	0.005	3.299	0.000
UMID.MED.4	0.021	0.006	3.296	0.000
UMID.MED.5	0.018	0.006	2.799	0.005
PRECIP.MED.1	0.002	0.005	4.336	$1.5e - 05$
PRECIP.MED.3	0.001	0.004	3.584	0.000
PRECIP.MED.5	0.002	0.000	3.949	$8.06e - 05$
PRECIP.MED.7	0.002	0.000	3.938	$8.44e - 05$
Tipo imóvel	-0.472	0.1089	-4.341	$1.47e - 05$
Quintal	0.305	0.094	3.235	0.001
Pequeno c	0.277	0.057	4.817	$1.54e - 06$
Suavização	<i>Edf</i>	Rank	F	<i>p</i> -valor
Coordenadas	24.69	29	5.253	$< 2e - 16$
Tempo	6.71	9	5.99	$2.35e - 08$

Tabela 15: Ajuste do modelo aditivo generalizado para o bairro Engenho do Meio.

O modelo mostra que os fatores de temperaturas 1 e 3 são altamente significativos, porém com sinas diferentes, a defasagem de 9 a 12 semanas indica que altas temperaturas são favoráveis ao mosquito, já a defasagem de 1 a 4 semanas, indica que a condição favorável é a de baixa temperatura. Para as covariáveis umidade e precipitação verifica-se que as semanas importantes estão entre 1 e 7, resultado que corrobora os encontrados para outros bairros, utilizando o fator umidade. Com relação as covariáveis das armadilhas, pode-se constatar a preferência do vetor por residências, sendo

que nestas localizações encontra-se em média 60.% (29.6% 98.6%) mais ovos que em não residenciais. O fato da localização da armadilha não apresentar quintal ou jardim foi considerado um fator de risco para este bairro, a ocorrência relativa aumenta em 35.7% (12.9% 63.2%) quando comparado as localizações que possuem quintal ou jardim. A presença de recipientes pequenos como jarros de barro ou baldes com tampa apresentou-se como um fator protetor, em localizações onde se encontra estes recipientes verifica-se em média 32% (18% 47.6%) a menos quando comparado as localizações que não apresentam tais recipientes. Este resultado merece uma análise mais detalhada que está fora do escopo deste trabalho, pode-se indicar como uma possível explicação para este resultado, a condição educacional, o fato do morador de determinada localização se preocupar em tampar jarros de barro ou baldes, pode refletir uma condição de maior informação e cuidado, afim de evitar criadouros do mosquito.

O efeito espacial mostrou-se altamente significativo, mostrando que a ocorrência de ovos do mosquito não é constante em toda a área o resultado da função suave das coordenadas geográficas das armadilhas pode ser visto na figura 16. Para a construção da superfície as covariáveis climáticas foram mantidas constantes na média, as covariáveis da armadilha na categoria de referência e o tempo em uma semana após a última observação.

O efeito temporal considerado como uma função suaves das datas de observação foi altamente significativo, pode-se observar um efeito sazonal bastante pronunciado, o período entre meados de fevereiro até final de junho são os meses de maior ocorrência de ovos, para os três anos observados. O que também confirma os resultados obtidos com as covariáveis referentes ao clima, em geral um período de altas temperaturas (outubro até começo de janeiro) são de baixa ocorrência, o mês de janeiro bastante quente porém com bastante chuva contribui para o aumento da série que chega no máximo sempre nos meses de maior umidade entre final de maio e junho, mostrando que os resultados do ajuste da função suave e das covariáveis climáticas são coerentes.

Para verificar a qualidade do ajuste foi realizado uma análise de resíduos através de técnicas gráficas, com base nos gráficos verifica-se que nenhum

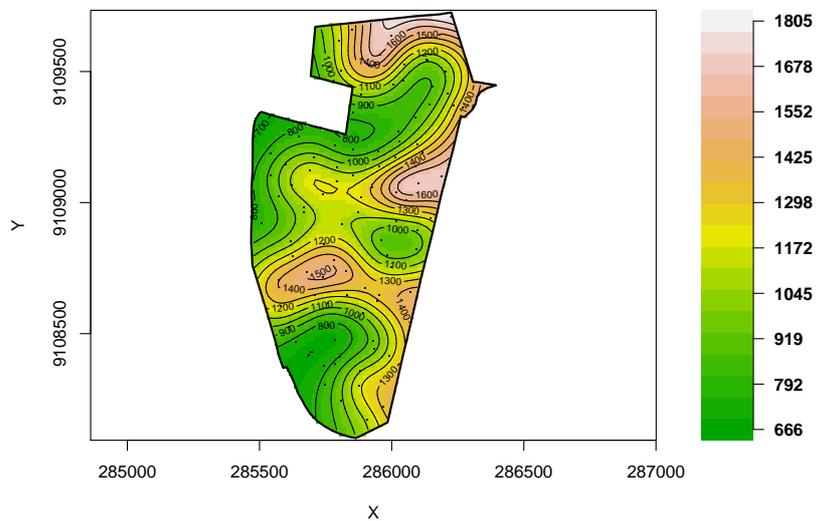


Figura 16: Superfície estimada da ocorrência de ovos de *Aedes aegypti* - Engenho do Meio.

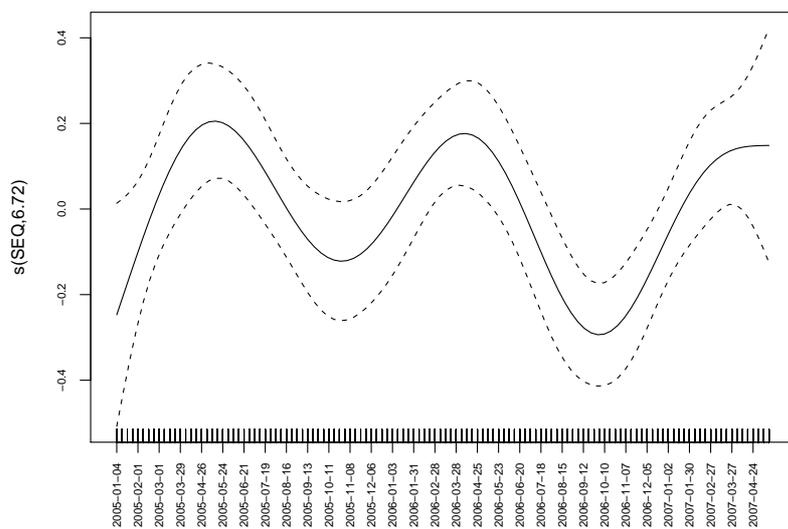


Figura 17: Ajuste do efeito temporal - Engenho do Meio.

dos pressupostos estão sendo violados, sendo assim, conclui-se que o modelo está bem ajustado

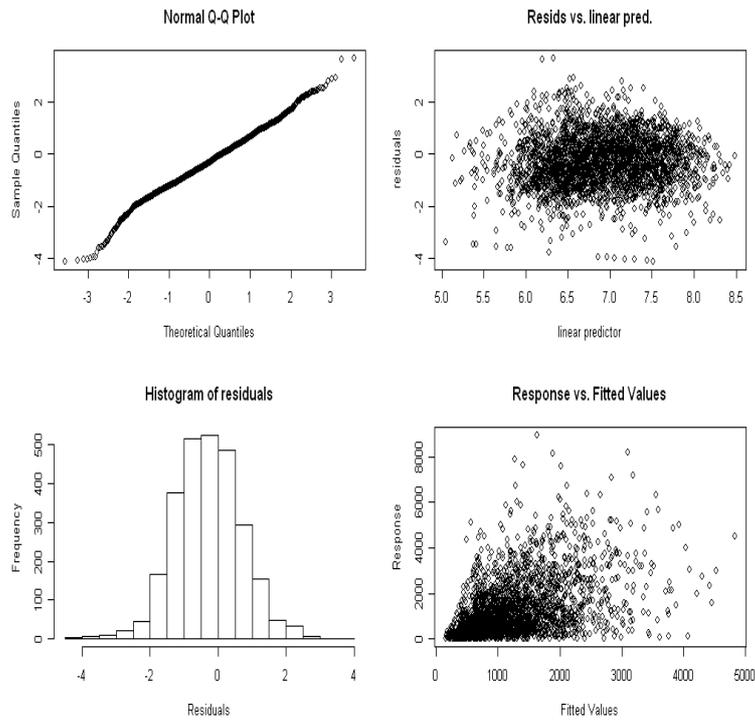


Figura 18: Análise dos resíduos do modelo com efeito espaço-temporal para o bairro Engenho do Meio.

5 Conclusão

A determinação de fatores associados à ocorrência de ovos de *Aedes aegypti*, é um processo delicado, que envolve conhecimentos em modelagem estatística e também sobre aspectos ligados ao comportamento biológico do mosquito. O uso de modelos estatísticos permite uma visão mais clara dos fatores relevantes por combinar as diversas covariáveis auxiliando na determinação precisa sobre as condições ambientais favoráveis ao aumento de contagens de ovos.

A análise da associação com fatores climáticos é uma tarefa difícil, já que, não se sabe a priori quais seriam as defasagens no tempo entre a data da coleta dos ovos e as condições climáticas que apresentam maior associação com tais contagens. Desta forma, o uso da análise fatorial foi de extrema importância por diminuir o número de covariáveis a serem analisadas e gerar vários cortes de defasagens que posteriormente puderam ser testados no modelo, evitando ainda a ocorrência de problemas ligados à multicolinearidade. Com relação aos fatores climáticos os resultados foram praticamente os mesmos para os cinco bairros em análise. De forma geral pode-se dizer que a condição climática mais favorável para a ocorrência de ovos de *Aedes* identificada pelos modelos resume-se em um período seco com altas temperaturas, um período de aproximadamente um mês para a reprodução, seguido de um período de alta umidade e queda na temperatura. Esta combinação mostrou-se bastante evidente nos bairros analisados.

A avaliação das covariáveis ligadas às armadilhas é de fundamental importância pois são elas que vão orientar as políticas de prevenção ao vetor. Com base nos resultados é possível confirmar o efeito das variáveis coletadas, as quais se supunham potencialmente associadas à contagens. Pode-se dizer que campanhas educacionais devem priorizar características com plantas em vasos, charco/poça, garrafas, fossa externa, piscina, poço elevador, laje sem telhado ou calhas que são características que atraem o vetor fazendo com que a ocorrência de ovos aumente significativamente. Outra característica importante detectada nas análises é a preferência do mosquito por residências.

A falta de abastecimento de água constante é um fator de risco e, sendo assim, programas governamentais devem ser implementados para levar esta condição a toda população.

Alguns resultados como a presença de recipientes pequenos sem tampa, indicado como um fator protetor, não eram esperados. Este resultado pode indicar que existe algum tipo de competição entre estes recipientes e as ovitrampas, fazendo com que o mosquito prefira por os ovos em jarros de barro ou baldes, diminuindo assim a ocorrência de ovos nas ovitrampas. Um resultado intrigante é o fato da localização da armadilha possuir água ligada a rede geral ser apontada como um fator de risco, esperava-se que nestas localizações a ocorrência de ovos fosse menor, pois são áreas já urbanizadas, com um mínimo de infra-estrutura.

A significância para os cinco bairros analisados do efeito espacial é muito importante, pois além de melhorar o ajuste do modelo estatístico permite verificar através de uma superfície onde estão as localizações de maior ocorrência de ovos. Isto é importante na prática, pois através de tais mapas as agências locais de saúde podem melhor direcionar suas ações de combate ao vetor. O fato deste efeito ser significativo também indica que podem haver outras covariáveis não levadas em consideração neste trabalho que devem ser importantes para explicar a resposta. Tais covariáveis teriam padrão espacial de distribuição na área e o efeito destas covariáveis omissas são então refletidos pela significância do efeito espacial.

O ajuste do efeito temporal permite identificar os meses do ano de maior ocorrência de ovos, além de indicar possíveis efeitos sazonais e verificar a tendência de crescimento ou decréscimo da série. De forma geral pode-se concluir que os modelos escolhidos apresentam um ajuste satisfatório, nenhuma quebra de pressupostos foi detectada, e portanto as conclusões são válidas.

O fato dos dados serem provenientes de um experimento real, onde o nível de controle sobre os diversos fatores que podem afetar a variável resposta é relativamente baixo, este estudo deve ser encarado como um experimento *screening* ou exploratório, que busca evidências de quais seriam os possíveis fatores que estão influenciando na resposta.

Outro aspecto importante a ser levantado, é a confiabilidade dos dados, já que o experimento é bastante grande requerendo o trabalho de muitas pessoas. A ocorrência de erros não aleatórios, como erros de digitação, a quantidade excessiva de dados faltantes, tanto das covariáveis como na resposta, a presença de valores fortemente discrepantes são alguns dos aspectos levantados neste trabalho e que certamente merecem revisão cuidadosa junto às fontes. Análises devem ser revistas em função das eventuais alterações nos dados.

Deixa-se aqui como agendas para pesquisas futuras relacionadas a este experimento, investigar o ajuste de um modelo assumindo para a variável resposta uma distribuição binomial negativa inflacionada de zeros. Também é recomendável uma aproximação da equipe de modelagem estatística com profissionais da área entomológica, preferencialmente profissionais da biologia. Recomenda-se uma análise detalhada das observações de cada uma das armadilhas junto a equipe de campo, para se identificar possíveis erros de digitação ou de natureza não aleatória, bem como a revisão dos resultados mediante diferentes tratamentos de dados atípicos e influentes.

6 Um olhar crítico aos dados.

Como já foi apontado na seção anterior uma das limitações deste trabalho é a veracidade dos dados. Optou-se por incluir esta nova seção para descrever alguns dos problemas encontrados, e que certamente limitaram o ajuste dos modelos. Cabe ressaltar que nenhuma das observações foi retirada do banco para o ajuste do modelo, esta decisão foi decorrente da equipe de campo garantir a veracidade dos dados.

O primeiro aspecto a ser levantado é referente as covariáveis das armadilhas. O experimento conta com 564 armadilhas nas quais as 10 covariáveis foram medidas apenas na data de implantação do experimento, sendo que nenhuma atualização após esta data foi realizada. Assumir que as mesmas condições foram mantidas desde a data de implantação do experimento é bastante forte, e isso pode ser uma das explicações de alguns resultados incoerentes quando foram avaliadas estas características.

A grande quantidade de dados faltantes para algumas armadilhas, é outro aspecto que influenciou no ajuste dos modelos. O conjunto de quatro gráficos apresentados da figura 19, mostra a quantidade excessiva de dados perdidos para esta armadilha, levando a pensar que uma armadilha como esta possa ser retirada da análise.

Um problema observado repetidamente nos gráficos de probabilidade normal e histograma dos resíduos, usados para verificar a qualidade do ajuste, foi a calda levemente mais pesada que a esperada pela distribuição Normal. Este fato decorre do excesso de zeros nas contagens em relação ao predito pelo modelo. Em algumas armadilhas observa-se que estes zeros ocorrem de forma inesperada ou suspeita onde a série de forma geral vem seguindo um padrão e repentinamente um zero ocorre, e logo depois deste zero, a série retorna com um pico de valor elevado. Este fato pode ser visto na figura 20, levando ao questionamento de que nesta data talvez a observação não tenha sido de fato realizada e que na próxima observação ocorreu um acúmulo de ovos ocasionando um pico na série.

O casos mostrados nesta seção são apenas alguns selecionados entre

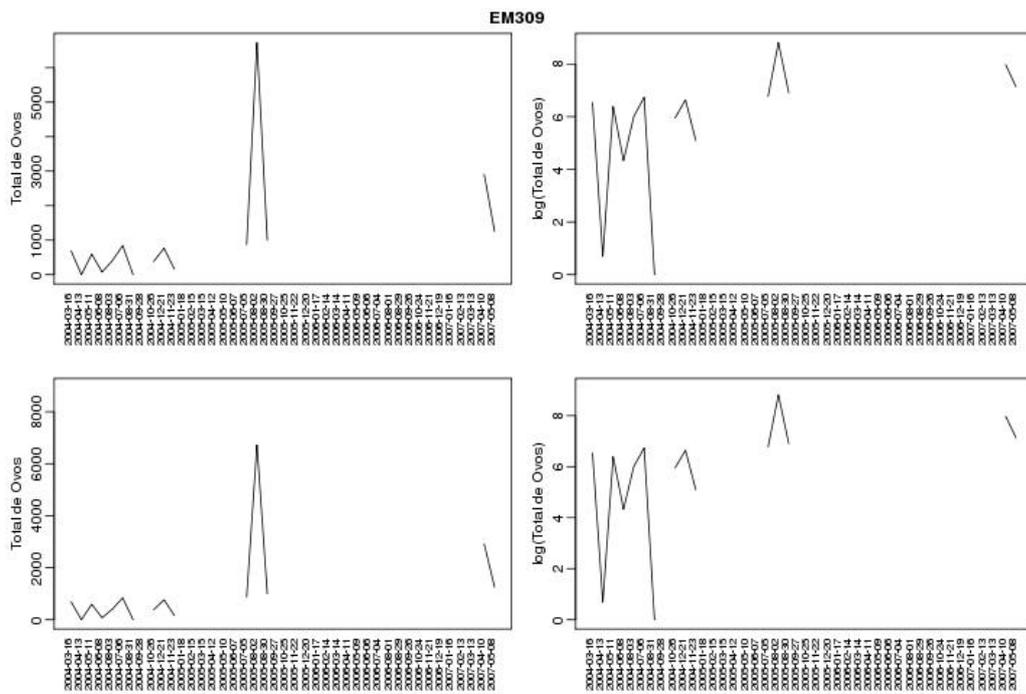


Figura 19: Observações da armadilha 09 do grupo 3, bairro Engenho do Meio.

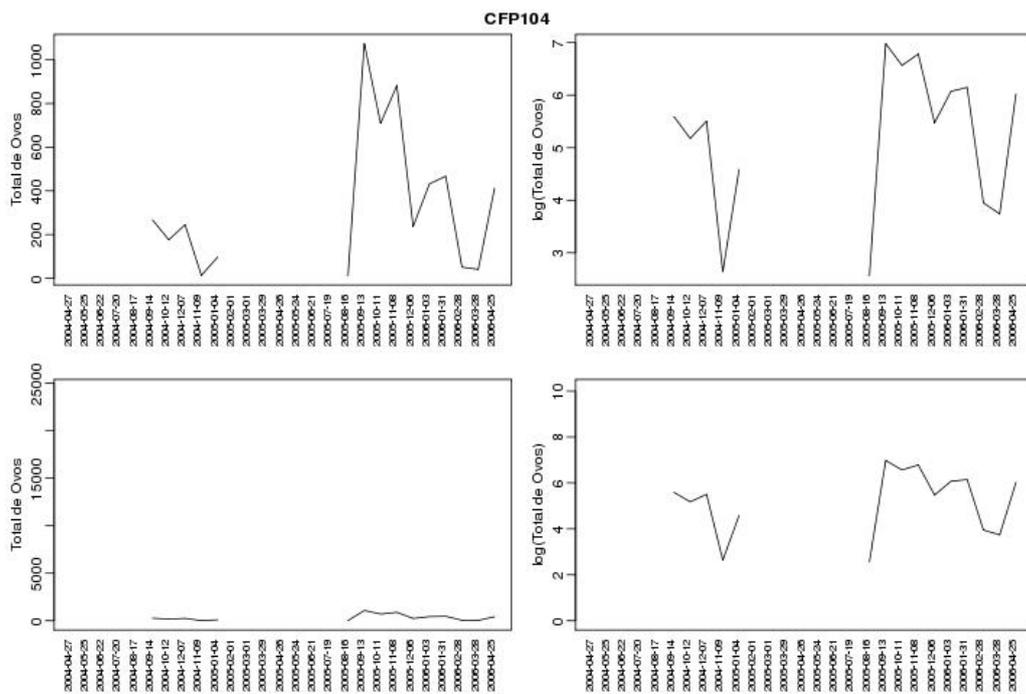


Figura 20: Observações da armadilha 04 do grupo 1, bairro Casa Forte/Parnamirim

vários outros comportamentos suspeitos que ocorrem no banco do SAU-DAVEL/Recife também detectados em outras armadilhas. Isto evidencia que as condições dos dados muitas vezes não são como se esperava, e é claro que estes casos influenciaram demasiadamente o ajuste dos modelos. Tais observações foram simplesmente tomadas como verdadeiras e os modelos ajustados incluindo tais observações. Esses fatos evidenciam que uma análise crítica dos dados deve ser feita junto à equipe de campo, para se confirmar e decidir como tratar caso a caso dados com comportamento atípico como os mostrados aqui.

Referências

- [Akaike, 1974] Akaike, H. 1974. A new look to statistical model identification. *IEEE Transactions on Automatic Control*, 716–722.
- [Braga & Valle, 2007] Braga, I. A., & Valle, D. 2007. *Aedes aegypti*: vigilância, monitoramento da resistência e alternativas de controle no Brasil. *Revista Brasileira de Epidemiologia*, **16**(4), 295–302.
- [Connor & Monroe, 1923] Connor, M. E., & Monroe, W. M. 1923. Stegomyia indices and their value in yellow fever control. *American Journal of tropical Medicine and Hygiene*, **3**, 9–19.
- [Dallazuanna *et al.*, 2008] Dallazuanna, H. S., Bonat, W. H., & Ribeiro Jr, P. J. 2008. RDengue um ambiente para monitoramento de ovos do mosquito *Aedes aegypti*. *Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria*.
- [Development Core Team, 2007] Development Core Team, R. 2007. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- [Fay & Eliason, 1966] Fay, R. W., & Eliason, D. A. 1966. A preferred oviposition site as a surveillance method for *Aedes aegypti*. *Mosquito News*, **26**, 531–534.
- [Fay & Perry, 1965] Fay, R. W., & Perry, A. S. 1965. Laboratory studies of ovipositional preferences of *Aedes aegypti*. *Mosquito News*, **25**, 270–281.
- [Ferreira & Chiaravalloti, 2007] Ferreira, A. C., & Chiaravalloti, F. 2007. Infestação de área urbana por *Aedes aegypti* e relação com níveis socioeconômicos. *Revista de Saúde Pública*, **41**(6), 15–22.
- [FNS, 2001] FNS. 2001. *Dengue - Instruções para o pessoal de combate ao vetor*. Fundação Nacional de Saúde, Brasília, DF.
- [Focks, 2000] Focks, D. A. 2000. A review of entomological sampling methods and indicators for dengue vectors. *Monografia na internet*.

- [Gomes, 2002] Gomes, A. C. 2002. Medidas dos níveis de infestação urbana para *Aedes (stegomyia) aegypti* e *Aedes (stegomyia) albopictus* em programas de vigilância entomológica. *Informe epidemiológico do SUS*, **11**(2), 79–90.
- [Gu, 2002] Gu, C. 2002. *Smoothing Splines ANOVA Models*. Springer.
- [Hastie & Tibishirani, 1990] Hastie, T. J., & Tibishirani, R. J. 1990. *Generalized Additive Models*. Chapman and Hall.
- [Johnson & Wichern, 1998] Johnson, R. A., & Wichern, D. W. 1998. *Applied multivariate statistical analysis*. 4 edn. New Jersey: Prentice Hall, Inc.
- [Lawley, 1942] Lawley, D. N. 1942. Further estimation in factor analysis. *Proceeding of the Royal Society of Edinburgh*.
- [McCullagh & Nelder, 1989] McCullagh, P., & Nelder, JA. 1989. *Generalized linear models*. 2 edn. London: Chapman and Hall.
- [Monteiro *et al.*, 2006] Monteiro, A. M., Carvalho, M. S., Assunção, R., Vieira, W., Ribeiro Jr, P. J., Davis Jr, C., & Régis, L. 2006. SAUDAVEL: Bridging the gap between Research and Services in Public Health Spatial Programs by Multi - Institucional Networking Development and Use of Spatial Information Technology Innovative Tools.
- [Nelson, 1995] Nelson, M. 1995. The significance of indicators and indices. *CTD/FIL/(DEN)/IC/95*.
- [OPS, 1995] OPS. 1995. *Dengue y dengue hemorrágico en las Américas: guías para su prevención y control*. Organización Panamericana de la Salud, Washington,DC.
- [Paula, 2004] Paula, G. A. 2004. *Modelos de Regressão com apoio computacional*. EdUSP Editora da Universidade de São Paulo.
- [Regis *et al.*, 2008] Regis, L., Monteiro, A. M., Melo Santos, M. A., Silveira, J. C., Furtado, A. F., Acioli, R. V., Santos, G. M., Nakasawa, M., Carvalho, M. S., Ribeiro Jr, P. J., & Souza, W. V. 2008. Developing new

- approaches for detecting and preventing *Aedes aegypti* population outbreaks: bases for surveillance, alert and control system. *Memórias do Instituto Oswaldo Cruz*, **103**, 50–59.
- [Ribeiro *et al.*, 2007] Ribeiro, P. A., Carreto, T., & Ribeiro Jr, P. J. 2007. *aRT:R-TERRALIB API*. R package version 1.5-0.
- [Santos, 1999] Santos, R. S. 1999. Fatores associados à ocorrência de formas imaturas de *Aedes aegypti* na Ilha do Governador, Rio de Janeiro, Brasil. *Revista da Sociedade Brasileira de Medicina Tropical*, **32**(4), 373–382.
- [Serpa *et al.*, 2007] Serpa, L. N., Costa, K. V., Voltolini, J. C., & Kakitani, I. 2007. Variação sazonal de *Aedes aegypti* e *Aedes albopictus* no município de Potim, São Paulo. *Revista de Saúde Pública*.
- [Silveira *et al.*, 2004] Silveira, J. C., Souza, W. V., Régis, L., Santos, M. A., Lapa, T. M., Portugal, J. L., Braga, T. S., & Monteiro, A. M. 2004. Recife em pedaços : Geotecnologias para a Detecção e Acompanhamento em Vigilância Epidemiológica. *VI Congresso Brasileiro de Epidemiologia*.
- [Tauil, 2002] Tauil, P. L. 2002. Aspectos críticos do controle do dengue no Brasil. *Cadernos de Saúde Pública*, **18**, 867–871.
- [Wahba, 2000] Wahba, G. 2000. *Smoothing Splines in Nonparametric Regression*. Tech. rept. University of Wisconsin.
- [Wood, 2008] Wood, S. 2008. *GAMs with GCV smoothness estimation and GAMMs by REML/PQL*. R package version 1.3-31.
- [Wood, 2006] Wood, S. N. 2006. *Generalized Additive Models: Introduction with R*. Chapman and Hall.
- [Yang & Thomé, 2007] Yang, M. H., & Thomé, R. C. A. 2007. Controle ótimo do Mosquito *Aedes aegypti* via técnica de insetos estéreis. *Congresso Nacional de Matemática Aplicada e Computacional*.

ANEXO