

ANA CAROLINA CINTRA NUNES MAFRA

**MODELAGEM MULTINOMIAL PARA A DISTRIBUIÇÃO
ESPACIAL DO RISCO EPIDEMIOLÓGICO**

Campinas

2011



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Ciências Médicas

**MODELAGEM MULTINOMIAL PARA A DISTRIBUIÇÃO
ESPACIAL DO RISCO EPIDEMIOLÓGICO**

Ana Carolina Cintra Nunes Mafra

Tese de Doutorado apresentada ao Programa de Pós-graduação em Saúde Coletiva do Departamento de Saúde Coletiva da Faculdade de Ciências Médicas da Universidade Estadual de Campinas para obtenção do título de Doutor em Saúde Coletiva, área de concentração Epidemiologia. Sob orientação do Prof. Dr. Ricardo Carlos Cordeiro.

Campinas, 2011

FICHA CATALOGRÁFICA ELABORADA POR
ROSANA EVANGELISTA PODEROSO – CRB8/6652
BIBLIOTECA DA FACULDADE DE CIÊNCIAS MÉDICAS
UNICAMP

M269m

Mafra, Ana Carolina Cintra Nunes, 1982
Modelagem multinomial para a distribuição espacial
do risco epidemiológico. / Ana Carolina Cintra Nunes
Mafra. -- Campinas, SP : [s.n.], 2011.

Orientador : Ricardo Carlos Cordeiro
Tese (Doutorado) - Universidade Estadual de
Campinas, Faculdade de Ciências Médicas.

1. Modelos logísticos. 2. Mapa de risco. 3. Risco
ajustado. 4. Estatísticas não-paramétricas. 5. Análise
espacial. I. Cordeiro, Ricardo. II. Universidade Estadual
de Campinas. Faculdade de Ciências Médicas. III.
Título.

Informações para Biblioteca Digital

Título em inglês: Multinomial models to estimate the spatial risk in epidemiology

Palavra-chave em inglês:

Logistic models

Risk map

Risk adjustment

Nonparametric statistics

Spatial analysis

Área de concentração: Epidemiologia

Titulação: Doutor em Saúde Coletiva

Banca examinadora:

Ricardo Carlos Cordeiro [Orientador]

Cláudia Torres Codeço

Júlio da Motta Singer

Djalma de Carvalho Moreira Filho

Ronaldo Dias

Data da defesa: 21-06-2011

Programa de Pós-Graduação: Faculdade de Ciências Médicas

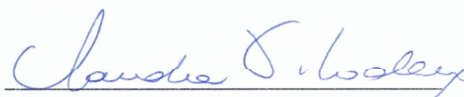
Banca examinadora de Tese de Doutorado

Aluno(a): Ana Carolina Cintra Nunes Mafra

Orientador: Prof.(a). Dr.(a). Ricardo Carlos Cordeiro



Prof.(a). Dr(a) Ricardo Carlos Cordeiro



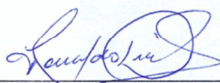
Prof.(a). Dr(a) Cláudia Torres Codeço



Prof.(a). Dr(a). Júlio da Motta Singer



Prof.(a). Dr(a). Djalma de Carvalho Moreira
Filho



Prof.(a). Dr(a). Ronaldo Dias

Curso de Pós-Graduação em Saúde Coletiva da Faculdade de Ciências Médicas da
Universidade Estadual de Campinas.

Data: 21/06/2011

Dedicatória

Toda obra de minhas mãos e intelecto dedico a Deus, que me deu a capacidade de pensar, investigar, criar e compreender o mundo que Ele criou.

Que esta obra contribua com todos que buscam fazer deste um mundo um pouco melhor.

Agradecimentos

Agradeço ao meu orientador, Ricardo Cordeiro, pelas oportunidades de crescimento, pelo incentivo e pelo apoio a todo instante.

À Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo suporte financeiro (processo 08/58461-5).

Aos professores e colegas do laboratório epiGeo, Liciane V. de A. Silveira, Luciana B. Nucci, Lia T. O. Zangirolani, Maria Rita D. Cordeiro, Celso Stephan, José Raimundo de S. Passos e todos os demais que muito contribuíram com ideias, estudos, trabalhos e se tornaram reais amigos durante esta caminhada.

Aos professores do Programa de Pós-Graduação em Saúde Coletiva por compartilhar conosco seu conhecimento e experiência.

Ao meu marido, Fabiano Lacerda Mafra, pelo companheirismo, pelo incentivo nos momentos difíceis e pela abnegação em me compartilhar com o doutorado e aceitar as consequências disso.

Aos meus pais, Sônia Maria C. Nunes e José Mauro Nunes, pelo esforço em nos proporcionar bom estudo, pela dedicação integral e por todo o encorajamento.

Aos amigos Ana Carolina G. S. Brianeze, Juliano R. Brianeze, Priscila Bratfisch, Milena S. F. Silva e todos da “República da ABU” por abrirem as portas de suas casas para me abrigar quando precisei me hospedar em Campinas em prol da pesquisa.

Às professoras que participaram de minha qualificação, Liciane V. de A. Silveira, Silvia E. Shimakura e Marília Sá Carvalho, pelas contribuições e apontamentos realizados naquela ocasião e aos professores que fizeram parte da banca de defesa desta tese.

E a todos os professores, funcionários, colegas, amigos e familiares que, apesar de não estarem nomeados, contribuíram para a realização deste projeto.

Resumo

A busca em compreender determinados fenômenos epidemiológicos muitas vezes envolve uma ferramenta denominada análise espacial do risco. O estudo do espaço em que ocorrem determinados desfechos permite ao pesquisador considerar informações não coletadas através de questionários ou prontuários médicos. Também insere questões sobre o que faz com que determinada área dentro da região de estudo se associe com maior risco ou proteção para o desfecho estudado. Existem muitos métodos para obter análises espaciais do risco, como os modelos aditivos generalizados, que permitem incluir nestas análises outras informações de interesse dos indivíduos estudados. Porém, atualmente, os estudos epidemiológicos que consideram a distribuição espacial do risco são analisados apenas com desfechos dicotômicos como, por exemplo, quando se classifica o indivíduo em doente ou não-doente. Esta é uma limitação que este trabalho visa superar ao apresentar um processo analítico da distribuição espacial do risco quando se tem uma variável resposta multinomial. Além de apresentar esta nova ferramenta, este trabalho analisou dois desfechos epidemiológicos: o primeiro é proveniente de um estudo caso-controle sobre acidentes de trabalho na cidade de Piracicaba em que a resposta foi: casos graves, casos leves ou controles; outra ilustração provém de um estudo transversal sobre criadouros de mosquitos no Distrito Sul de Campinas, onde se encontrou muitos criadouros, poucos criadouros ou nenhum criadouro. Primeiramente, faz-se necessária uma discussão sobre a adequação de cada modelo multinomial a alguns estudos epidemiológicos. Também se discute a escolha de um entre diversos modelos multinomiais e apresenta-se a maneira de interpretar os resultados da análise. Para tornar este método acessível a outros pesquisadores, são apresentadas funções computacionais para o processo analítico.

Palavras-chave: Modelos multinomiais; Mapa de risco; Risco ajustado; Modelos logísticos; Estatística não-paramétrica.

Abstract

The search for understanding some epidemiological phenomena often involves an tool called spatial analysis of risk. The study of space in which certain outcomes occur allows the researcher to consider information that can not be collected through questionnaires or medical records. It also puts questions about what makes a certain area within the study region was associated with greater risk or protection for the outcome studied. Many techniques are used for this kind of study as the generalized additive models that fit the spatial analysis of the risk with others informations of interest. But now, epidemiological studies that consider the spatial distribution of risk are analyzed only with dichotomous outcomes, such as when it classifies the individual in case or control. This is a limitation that this study aims to overcome when presenting an analytical process of the spatial distribution of risk when you have a multinomial response variable. In addition to presenting this new tool, this study analyzed two outcomes: first, from a case-control study of precarious workers in the city of Piracicaba in which the response was: severe cases, mild cases or controls. Another illustration comes from a cross-sectional study on mosquito breeding sites in the Southern District of Campinas, where we met many breeding sites, few or no breeding sites. First, it is necessary a discussion on the appropriateness of each multinomial model to some epidemiological studies. It also discusses the choice of one among several multinomial models and shows the way to interpret the results of the analysis. We present the computational functions for the analytical process to make this method accessible to other researchers.

Keywords: Multinomial models; Risk map; Risk adjustment; Logistic models; Nonparametric statistics.

Lista de Ilustrações

Quadros

| | |
|---|----|
| Quadro 1: Funções utilizadas no algoritmo para obtenção dos GAM semiparamétricos com respostas multinomiais. | 25 |
| Quadro 2: Análises espaciais multinomiais com respectivas funções de risco espacial e modelo logístico. | 35 |

Tabelas

| | |
|---|----|
| Tabela 1: Médias e desvios padrão das variáveis contínuas e porcentagens de indivíduos em cada classe por categorias de resposta. Acidentes de trabalho em Piracicaba, 2007. | 44 |
| Tabela 2: Razões de chances (RC) obtidas através do ajuste do modelo binomial dos dados de acidente de trabalho em Piracicaba, 2007 (N=2323). | 47 |
| Tabela 3: Razões de chances (RC) obtidas através do ajuste do modelo multinomial dos dados de acidente de trabalho em Piracicaba, 2007 (N=2323). ... | 47 |
| Tabela 4: Médias e desvios padrões da escolaridade do chefe de domicílio e porcentagens de indivíduos em cada classe por categorias de resposta. Criadouros de mosquitos no Distrito Sul de Campinas, 2007. | 55 |
| Tabela 5: Razões de chances obtidas através do ajuste dos modelos linear generalizado (paramétrico) e aditivo generalizado (semiparamétrico) para dados de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007 (N=727). | 57 |

Figuras

| | |
|---|----|
| Figura 1: Distribuição espacial dos locais de acidente e trabalho em Piracicaba, 2007. A) Pontos dos casos leves e dos controles. B) Pontos dos casos graves e controles. | 45 |
| Figura 2: Distribuição espacial do risco relativo (RR) de acidentes de trabalho em Piracicaba, 2007. Risco de acidentes de trabalho (modelo binomial) bruto (A) e ajustado (B). Risco de casos leves em relação aos controles (modelo multinomial) bruto (C) e ajustado (D). Risco de casos graves em relação aos controles (modelo multinomial) bruto (E) e ajustado (F). | 50 |
| Figura 3: Distribuição do número de possíveis criadouros encontrados no Distrito Sul de Campinas, 2007. | 54 |
| Figura 4: Distribuição espacial dos domicílios estudados no Distrito Sul de Campinas, 2007. A) Pontos onde não se encontrou nenhum criadouro. B) Pontos em que se encontrou de um a dez criadouros. C) Pontos onde se encontrou mais de dez criadouros. | 55 |

| | |
|--|----|
| Figura 5: Distribuição espacial do risco relativo (RR) ajustado de criadouros de mosquitos em potencial no Distrito Sul de Campinas, 2007. Modelos multinomiais cumulativo (A e B), de razões sequenciais (C e D), de categorias adjacentes (E e F), politômico (G e H) e modelo binomial (I)..... | 59 |
| Figura 6: Análise dos resíduos do modelo binomial apenas com informação espacial dos acidentes de trabalho em Piracicaba, 2007..... | 80 |
| Figura 7: Análise dos resíduos do modelo binomial semiparamétrico dos acidentes de trabalho em Piracicaba, 2007..... | 80 |
| Figura 8: Análise dos resíduos do modelo multinomial apenas com informação espacial dos acidentes de trabalho em Piracicaba, 2007. A) Logito entre acidentes leves e controles. B) Logito entre acidentes graves e controles..... | 81 |
| Figura 9: Análise dos resíduos do modelo multinomial semiparamétrico dos acidentes de trabalho em Piracicaba, 2007. A) Logito entre acidentes leves e controles. B) Logito entre acidentes graves e controles..... | 82 |
| Figura 10: Análise dos resíduos do modelo binomial semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007. | 83 |
| Figura 11: Análise dos resíduos do modelo cumulativo semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007. | 84 |
| Figura 12: Análise dos resíduos do modelo de razões sequenciais semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007. | 85 |
| Figura 13: Análise dos resíduos do modelo de categorias adjacentes semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007. | 86 |
| Figura 14: Análise dos resíduos do modelo politômico semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007. | 87 |

Lista de abreviaturas e siglas

AIC: Critério de Informação de Akaike ou *Akaike's Information Criterion*.

CV: validação cruzada ou *cross-validation*.

GAM: modelo aditivo generalizado ou *generalized additive model*.

GLM: modelo linear generalizado ou *generalized linear model*.

LI: limite inferior de intervalo de confiança.

LS: limite superior de intervalo de confiança.

RC: razão de chances ou *odds ratio*.

Lista de notações

\wedge : indica que determinado valor é estimado;

r : número de categorias da variável resposta multinomial;

n : número de indivíduos na amostra;

\ln : logarítimo natural;

i : indicador de indivíduos; varia de 1 a n ;

j : indicador de categoria ou de função; varia de 1 a r ;

$\lambda(\cdot)$: função densidade espacial de ocorrência de um determinado evento;

q : fração amostral, também chamada de proporção de indivíduos na amostra em relação ao total de indivíduos na população;

\mathfrak{R} : região de estudo;

\mathbf{x} : matriz de informações espaciais, ou seja, latitude e longitude dos pontos no estudo;

\mathbf{u} : matrix de informações não-espaciais, ou seja, das demais variáveis independentes no estudo;

p : probabilidade de um determinado evento;

μ : valor esperado;

V : variância;

c : constante não conhecida;

$P(\cdot)$: função de probabilidade;

$E(\cdot)$: função esperança;

ρ : log-risco espacial;

θ : risco espacial;

Θ : risco relativo espacial;

η : preditor linear;

α : intercepto de modelos;

β ou $\boldsymbol{\beta}$: parâmetros estimados por um modelo ou matriz/vetor destes;

b : número de parâmetros de um modelo;

Y, y : denotam a variável resposta e os valores que esta pode assumir;

$f(\cdot)$: função de ligação;

$g(\cdot)$: função não-paramétrica;
 z : variável dependente ajustada;
 w : pesos;
 h : tamanho de banda;
 $K_h(\cdot)$: Kernel Nadaraya-Watson estimado com banda h ;
 $\phi(\cdot)$: valor exponencial de um determinado logito;
 ∂ : derivada;
 Δ : indica variação;
 e : resíduo;
 ε : erro;
 d : resíduo *deviance*;
 D : *deviance* residual.

Sumário

| | |
|---|------------|
| Resumo | vi |
| Abstract..... | vii |
| 1. Introdução..... | 14 |
| 2. Objetivos | 17 |
| 3. Métodos..... | 18 |
| 3.1 Modelos estatísticos | 18 |
| 3.1.1 Modelos lineares generalizados | 19 |
| 3.1.2 Modelos aditivos generalizados | 20 |
| 3.1.3 Funções Logito..... | 21 |
| 3.1.4 GAM multinomiais | 23 |
| 3.2 Análise de variáveis não-espaciais | 28 |
| 3.3 Análise espacial do risco para dados com resposta binomial | 29 |
| 3.4 Análise espacial do risco para dados com resposta multinomial | 31 |
| 3.4.1 Definição do risco espacial multinomial..... | 32 |
| 3.4.2 Risco relativo espacial multinomial | 36 |
| 3.4.3 Intervalo de confiança para o risco relativo espacial | 40 |
| 3.5 Aspectos computacionais | 41 |
| 4. Estudo caso-controle com resposta multinomial | 42 |
| 4.1 Acidentes de trabalho em Piracicaba, SP..... | 43 |
| 4.2 Resultados dos acidentes de trabalho | 45 |
| Análises espaciais..... | 48 |
| 4.3 Discussão sobre os resultados dos acidentes de trabalho | 51 |
| 5. Estudo transversal com resposta multinomial | 53 |
| 5.1 Criadouros de mosquito no Distrito Sul de Campinas, SP..... | 53 |
| 5.2 Resultados dos criadouros de mosquitos | 56 |
| Análises espaciais..... | 58 |
| 5.3 Discussão sobre os resultados dos criadouros de mosquitos | 60 |
| 6. Discussão Geral | 62 |
| 7. Conclusão | 64 |
| 8. Referências Bibliográficas | 65 |
| 9. Anexo: roteiro de identificação de trabalhado precarizado | 70 |
| 10. Apêndices | 71 |

| | |
|--|----|
| 10.1 Obtenção dos valores de p , z e w para o ajuste dos Modelos Aditivos Generalizados com resposta multinomial | 71 |
| 10.2 Solução para obter o risco relativo espacial na análise cumulativa. | 77 |
| 10.3 Análises de resíduos dos acidentes de trabalho..... | 79 |
| 10.4 Análises de resíduos dos criadouros de mosquitos. | 83 |
| 10.5 Referências de artigos relacionados a esta tese. | 88 |

1. Introdução

Na epidemiologia define-se o espaço como a “totalidade de coisas materiais (vivas e inanimadas), de cuja associação o mundo, ou parte dele, é formado”⁽¹⁾. Neste aspecto podemos considerar que determinados desfechos estudados na área da saúde podem ter associação não diretamente com o espaço, mas com o que o constitui, sejam aspectos geográficos, geopolíticos, político-administrativos, sócio-econômicos, que nem sempre conseguimos explicar por meio de variáveis específicas, ou que não conseguimos coletá-las de forma eficiente.

Desde Hipócrates (século V a.C.) a relação entre o espaço e a ocorrência de doenças é estudada⁽²⁾. John Snow, em seu famoso estudo da cólera em Londres no ano de 1854, ao analisar espacialmente os casos de cólera percebeu uma associação entre o local de moradia dos contaminados e algumas bombas de água na cidade de Londres, formulando a hipótese, posteriormente confirmada, de que a transmissão da doença se dava por meio hídrico⁽²⁾.

Atualmente com o desenvolvimento computacional existem muitas ferramentas para análise espacial de dados. Uma delas considera o ajuste de modelos aditivos generalizados (GAM), pois ele permite incorporar ao modelo covariáveis além do espaço, possibilitando a verificação do risco no espaço já controlado por tais covariáveis se estas forem significativas ao modelo. Ele também nos dá estatísticas que permitem que sejam feitas inferências, além de mapas de distribuição espacial do risco⁽³⁾.

A análise espacial do risco por meio de GAM é muito utilizada na Epidemiologia em análise de dados provenientes de estudos caso-controle⁽²⁾, como, por exemplo, os estudos dos autores Ballester et al.⁽⁴⁾, Bayentin et al.⁽⁵⁾, Braga et al.⁽⁶⁾, Carvalho et al.⁽⁷⁾, French e Wand⁽⁸⁾, Honório et al.⁽⁹⁾, Ozonoff et al.⁽¹⁰⁾, Shimakura et al.⁽¹¹⁾, Stephan et al.⁽¹²⁾, Vieira et al.⁽¹³⁾, Vieira et al.⁽¹⁴⁾, Vieira et al.⁽¹⁵⁾, Webster et al.⁽¹⁶⁾, Young et al.⁽¹⁷⁾ e Zangirolani et al.⁽¹⁸⁾.

Em decorrência dos estudos caso-controle classificarem os indivíduos estudados como doentes ou não doentes, neles, muito frequentemente, a relação entre exposição e adoecimento é modelada pela distribuição binomial, por meio de ajustes de modelos lineares generalizados (GLM), tendo como função de ligação o

logito da probabilidade de doença. Entretanto, nem sempre é conveniente classificar dicotomicamente a população estudada. Pode ser interessante, para os propósitos do estudo, subdividir os casos analisados de acordo com características tais como gravidade, tipo celular, forma clínica, etc., obtendo-se maior compreensão do desfecho e sua relação com as exposições estudadas. Essa subdivisão resulta em subconjuntos dos indivíduos estudados, formados pelos diferentes tipos de casos e pelos controles.

Subdividir o desfecho em mais de duas classes é uma opção que enriquece a análise dos dados, pois permite incorporar mais informação ao estudo dos fatores de risco associados ao agravo investigado e, por isso, muitos estudos epidemiológicos fazem uso desta abordagem⁽¹⁹⁻³¹⁾.

Entretanto nenhum dos estudos caso-controle que realizaram análises espaciais consideraram essa possibilidade e não foi encontrada na literatura ferramenta analítica que permitisse mesclar o estudo do espaço de ocorrência de doenças a fenômenos que possuem mais do que duas categorias de resposta, sendo este um dos objetivos desse trabalho.

Quando uma variável resposta possui mais do que duas categorias possíveis considera-se que esta possui uma distribuição de probabilidade chamada multinomial, uma extensão da distribuição binomial⁽³²⁾. Na análise deste tipo de dados existe uma distinção importante entre as abordagens nominal e ordinal.

A abordagem ordinal pressupõe uma ordem ou uma gradação entre as categorias do fenômeno estudado^(19, 33). São exemplos de modelos multinomiais ordinais os modelos de Categorias Adjacentes⁽³²⁾, de Razões Sequenciais⁽³⁴⁾, Cumulativo⁽³²⁾, entre outros. O modelo logístico de Categorias Adjacentes compara entre si duas categorias adjacentes ao longo de uma escala ordinal. O modelo de Razões Sequenciais compara entre si uma categoria com todas as outras superiores na escala ordinal. O modelo Cumulativo compara entre si categorias abaixo e acima de pontos de corte ao longo de uma escala ordinal.

Na abordagem nominal o modelo mais comum é o logístico Político⁽³⁵⁾ onde, escolhendo-se uma das categorias como referência, compara-se esta, um a

um, com todas as outras categorias. Aliás, este modelo não possui restrição quanto à ordinalidade da resposta, podendo ser usado também quando a resposta é ordinal⁽³³⁾.

Quanto à escolha entre qual abordagem multinomial seria mais adequada para a análise de dados provenientes de estudos caso-controle, uma revisão de literatura que considera estudos em que os casos foram subdivididos em dois ou mais grupos por tipo ou gravidade do caso aponta que, “apesar de grande parte dos artigos terem optado por utilizar em suas análises o modelo politômico (48%), não há na literatura discussão sobre essa escolha e sobre a adequação destas análises ao delineamento dos estudos caso-controle e aos pressupostos epidemiológicos relacionados a este”⁽³⁶⁾, algo que esse trabalho visou atender e ampliar essa discussão para outros tipos de estudos epidemiológicos.

Os objetivos dessa tese estão mais bem definidos no Capítulo 2.

O Capítulo 3 traz uma revisão sobre os modelos estatísticos utilizados para a análise espacial de dados e informações necessárias para a compreensão do processo analítico para análise espacial de dados com resposta multinomial, o qual foi desenvolvido nesse projeto de pesquisa e está apresentado na seção 3.4. Ainda apresenta as medidas utilizadas para comparar o novo método com os métodos analíticos usuais.

O novo método foi aplicado a dados provenientes de dois tipos de estudos epidemiológicos: caso-controle e transversal. Estes estão apresentados, analisados e discutidos nos Capítulos 4 e 5, respectivamente.

A discussão e a conclusão geral dessa tese encontram-se nos Capítulos 6 e 7.

2. Objetivos

Objetivo geral:

Possibilitar uma análise da distribuição espacial do risco quando o evento estudado está subdividido em mais do que duas classes.

Objetivos específicos:

1. Organizar um processo de análise que conceda o estudo da distribuição espacial do risco quando se tem desfechos multinomiais.

Para tanto, é preciso escolher a metodologia estatística que melhor auxilia na compreensão destes fenômenos e caso seja preciso, desenvolver o que ainda não estiver na literatura. A partir daí, definir o risco espacial multinomial (medida epidemiológica) com base nos modelos estatísticos utilizados e obter informações sobre a significância das estimativas. Por fim, é preciso desenvolver funções computacionais para a realização do processo analítico.

2. Analisar dados provenientes de estudos epidemiológicos em que o desfecho possui mais do que duas categorias de resposta.

3. Discutir a adequação do uso do método proposto para diferentes hipóteses e tipos de estudos epidemiológicos.

3. Métodos

O método de análise espacial do risco para dados com variável resposta com distribuição multinomial é um produto deste doutorado e, por ser um novo processo analítico, será comparado com as análises usuais da epidemiologia.

Uma análise usual que poderia ser feita nos dados desse trabalho seria estudar apenas as covariáveis não-espaciais (sem a inclusão da informação da localização geográfica dos indivíduos do estudo), visto que existem estudos epidemiológicos com desfechos multinomiais que não consideram a análise espacial do risco, como já comentado na introdução dessa tese. Os métodos utilizados neste tipo de análise estão apresentados na seção 3.2.

Outra análise usual a ser considerada é o estudo espacial do risco em que a variável resposta segue uma distribuição binomial. Essa análise é apresentada na seção 3.3.

A seção 3.4 traz o novo processo analítico para análise espacial do risco em dados com resposta multinomial, bem como as definições de risco espacial para este tipo de análise.

A seção 3.1 apresenta os modelos estatísticos utilizados nas análises usuais e também no novo processo analítico.

As informações a respeito das funções computacionais e dos *softwares* utilizados para todas as análises apresentadas encontram-se na seção 3.5.

3.1 Modelos estatísticos

Inicia-se este capítulo com esta seção para melhor entendimento das ferramentas que foram utilizadas nos processos de análise que serão apresentados nas demais seções.

As seções 3.1.1-3.1.3 trazem apenas uma revisão de literatura para introduzir conceitos utilizados para o desenvolvimento do processo analítico do risco espacial em casos que se tem uma variável resposta com distribuição multinomial e também para a análise dos dados nas maneiras usuais já citadas anteriormente.

O conteúdo da seção 3.1.4 é um produto deste doutorado para viabilizar o processo analítico proposto na seção 3.4.

3.1.1 Modelos lineares generalizados

É possível utilizar métodos análogos àqueles desenvolvidos para o modelo de regressão linear, em situações em que a variável resposta obedece a outras distribuições que não a Normal. Isto se deve, em parte, ao conhecimento de que muitas das propriedades da distribuição Normal são compartilhadas por uma larga classe de distribuições chamada de família exponencial de distribuições. Muitas distribuições conhecidas pertencem a essa família, como a própria Normal, a Poisson, a Binomial, a Multinomial, a Gama, a Binomial Negativa e a Gaussiana Inversa.

Os GLM foram primeiramente apresentados por Nelder e Wedderburn⁽³⁷⁾ e englobam os modelos de regressão linear simples e múltipla, regressão logística, regressão de Poisson e muitos outros. Os ingredientes básicos desses modelos são:

1. N valores independentes Y_1, \dots, Y_N , de uma variável resposta que segue uma distribuição da família exponencial, com valor esperado $E(Y_i) = \mu_i$;

2. N vetores $\mathbf{u}_i = (1 \quad u_{i1} \quad u_{i2} \quad \dots \quad u_{ib})^t$, $i=1, \dots, N$, contendo os valores de b variáveis explicativas;

3. uma função monotônica e diferenciável f , chamada de função de ligação, tal que $f(\mu_i) = \mathbf{u}_i^t \boldsymbol{\beta}$, $i=1, \dots, N$ (1)

com $\boldsymbol{\beta} = (\beta_1 \quad \beta_2 \quad \dots \quad \beta_b)$ representando o vetor de parâmetros a serem estimados.

O vetor de parâmetros $\boldsymbol{\beta}$ na equação (1) pode ser estimado pelo método de máxima verossimilhança, e os cálculos envolvem um procedimento iterativo⁽³⁷⁾.

É sobre a análise destes parâmetros que trata a seção 3.2, porém para as análises espaciais do risco é necessário outro modelo que permita incorporar a informação das localizações geográficas dos indivíduos como superfícies e, com este fim, usou-se os GAM.

3.1.2 Modelos aditivos generalizados

Os GAM⁽³⁾ são uma extensão dos GLM, em que, ao invés de incluir a localização espacial no modelo como as demais variáveis explicativas, ou seja, como $\mathbf{u}'\boldsymbol{\beta}_1 + \mathbf{x}'\boldsymbol{\beta}_2$, inclui uma função de \mathbf{x} da forma $\mathbf{u}'\boldsymbol{\beta} + g(\mathbf{x})$, sendo $g(\mathbf{x})$ uma função não-paramétrica (i.e. cuja forma não é especificada através de parâmetros) e que pode ser estimada através de curvas de alisamento⁽³⁾.

Com essa substituição, não é necessário assumir uma relação linear entre $f(\mu_i)$ e $g(\mathbf{x})$, como no GLM. De fato, não é necessário nem mesmo conhecer previamente a forma dessa relação, mas é possível estimá-la a partir de um conjunto de dados. Essa função estimada $g(\mathbf{x})$, também chamada de curva alisada ou *smoother*, em muitas situações, nada mais é do que uma média ponderada dos valores Y_i na vizinhança de um dado valor x_i .

A superfície alisada permite então descrever a forma, e mesmo revelar possíveis não-linearidades nas relações estudadas, uma vez que não apresenta a estrutura rígida de uma função paramétrica.

Hastie e Tibshirani⁽³⁾ apresentam uma série de alisadores, alguns mais refinados do que a média móvel, como os *cubic smoothing splines* ou o *locally weighted running line smoother (loess)*.

As duas considerações mais importantes para a escolha do alisador envolvem o tipo de função dos valores de y a ser estimada e de um parâmetro de suavização, que pode ser o tamanho da vizinhança, dependendo da função escolhida para a suavização. Este parâmetro determina a relação entre o viés e a variância do alisador.

A função de alisamento utilizada nas análises desse trabalho foi o Kernel de Nadaraya-Watson, também utilizado para estimação do risco espacial por Kelsall e Diggle⁽³⁸⁾.

Esses modelos foram utilizados para a análise espacial do risco quando assumimos que a variável resposta segue uma distribuição binomial (seção 3.3) e quando assumimos que segue uma distribuição multinomial (seção 3.4).

Isso porque, assim como os GLM, os GAM podem ser utilizados para análise de dados que seguem qualquer uma das distribuições de probabilidade

pertencentes à família exponencial, variando-se apenas a função de ligação, que, no caso das distribuições binomial e multinomial pode ser uma função logito.

3.1.3 Funções Logito

Por definição uma função logito é o logaritmo natural (\ln) da razão de duas probabilidades. Ela é muito utilizada na epidemiologia, tanto para dados com resposta binomial quanto multinomial, em razão do exponencial desta função prover a estimação de Razão de Chances (RC), quando utilizada num modelo de regressão como função de ligação.

No caso binomial são consideradas apenas duas probabilidades porque só existem dois eventos possíveis, então, se Y é a variável resposta e assume valores $y=1,2$, podemos chamar a probabilidade de cada um destes eventos de p_1 e p_2 , respectivamente. Neste caso a função logito é expressa assim:

$$\ln \left[\frac{P(Y=1)}{P(Y=2)} \right] = \ln \left[\frac{p_1}{p_2} \right]$$

Para utilizar esta função como função de ligação para o ajuste de um GLM em que β é o vetor de parâmetros associados às covariáveis \mathbf{u} e α é o intercepto, o modelo fica:

$$\ln \left[\frac{P(Y=1|\mathbf{u})}{P(Y=2|\mathbf{u})} \right] = \ln \left[\frac{p_1(\mathbf{u})}{p_2(\mathbf{u})} \right] = \alpha - \mathbf{u}'\beta$$

Se, ao invés de um GLM, for ajustado um GAM que, além das covariáveis \mathbf{u} considere a função não-paramétrica $g(\mathbf{x})$ da coordenadas geográficas \mathbf{x} , o modelo é expresso da seguinte maneira:

$$\ln \left[\frac{P(Y=1|\mathbf{x},\mathbf{u})}{P(Y=2|\mathbf{x},\mathbf{u})} \right] = \ln \left[\frac{p_1(\mathbf{x},\mathbf{u})}{p_2(\mathbf{x},\mathbf{u})} \right] = \alpha - \mathbf{u}'\beta + g(\mathbf{x})$$

Quando se tem uma variável resposta multinomial, a variável resposta Y possui mais de duas categorias – assume os valores $y=1,2,\dots,r$ sendo r o número de eventos possíveis. Neste caso, tem-se uma probabilidade associada a cada um dos r eventos (p_1, p_2, \dots, p_r) e para estruturar a função logito existem diversas maneiras de combinar estas probabilidades.

Uma maneira de estruturar o logito para dados multinomiais é eleger uma das categorias de resposta e compara-la com as demais, duas a duas. Este logito recebe o nome de logito de referência⁽³⁹⁾ e os modelos que fazem uso deste logito podem ser chamados de *baseline*⁽³²⁾ ou modelo logístico politômico⁽³³⁾.

Seguindo a mesma notação utilizada para os logitos binomiais, os modelos logísticos politômicos são expressos da seguinte maneira:

$$\text{GLM: } \ln \left[\frac{P(Y = y_j | \mathbf{u})}{P(Y = y_r | \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{u})}{p_r(\mathbf{u})} \right] = \alpha_j + \mathbf{u}'\boldsymbol{\beta}_j, \quad j = 1, 2, \dots, r-1$$

$$\text{GAM: } \ln \left[\frac{P(Y = y_j | \mathbf{x}, \mathbf{u})}{P(Y = y_r | \mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{x}, \mathbf{u})}{p_r(\mathbf{x}, \mathbf{u})} \right] = \alpha_j + \mathbf{u}'\boldsymbol{\beta}_j + g_j(\mathbf{x}), \quad j = 1, 2, \dots, r-1$$

Outra maneira de estruturar os logitos é comparar cada probabilidade com a sua adjacente. Os modelos que utilizam estes logitos são chamados modelos logísticos de categorias adjacentes⁽³³⁾ e podem ser expressos assim:

$$\text{GLM: } \ln \left[\frac{P(Y = y_j | \mathbf{u})}{P(Y = y_{j+1} | \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{u})}{p_{j+1}(\mathbf{u})} \right] = \alpha_j - \mathbf{u}'\boldsymbol{\beta}_j, \quad j = 1, 2, \dots, r-1$$

$$\text{GAM: } \ln \left[\frac{P(Y = y_j | \mathbf{x}, \mathbf{u})}{P(Y = y_{j+1} | \mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{x}, \mathbf{u})}{p_{j+1}(\mathbf{x}, \mathbf{u})} \right] = \alpha_j - \mathbf{u}'\boldsymbol{\beta}_j + g_j(\mathbf{x}), \quad j = 1, 2, \dots, r-1$$

Além de comparar probabilidades duas a duas, pode-se também soma-las.

Existem dois logitos que estruturam as probabilidades de forma a soma-las. Um deles, utiliza somas de probabilidades no numerador e no denominador das razões (logitos cumulativos), outro, apenas as soma no denominador (logitos sequenciais).

O modelo que utiliza logitos cumulativos é denominado modelo logístico cumulativo⁽³³⁾ e pode ser expresso da seguinte maneira:

$$\text{GLM: } \ln \left[\frac{P(Y \leq y_j | \mathbf{u})}{P(Y > y_j | \mathbf{u})} \right] = \ln \left[\frac{\sum_{l=1}^j p_l(\mathbf{u})}{\sum_{m=j+1}^r p_m(\mathbf{u})} \right] = \alpha_j - \mathbf{u}'\boldsymbol{\beta}_j, \quad j = 1, 2, \dots, r-1$$

$$\text{GAM: } \ln \left[\frac{P(Y \leq y_j | \mathbf{x}, \mathbf{u})}{P(Y > y_j | \mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{\sum_{l=1}^j p_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r p_m(\mathbf{x}, \mathbf{u})} \right] = \alpha_j - \mathbf{u}'\boldsymbol{\beta}_j + g_j(\mathbf{x}), \quad j = 1, 2, \dots, r-1$$

O modelo que utiliza logitos sequenciais é o modelo logísticos de razões sequenciais, em inglês chamado de *continuation-ratio model*⁽³³⁾ e é expresso da seguinte maneira:

$$\text{GLM: } \ln \left[\frac{P(Y = y_j | \mathbf{u})}{P(Y > y_j | \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{u})}{\sum_{m=j+1}^r p_m(\mathbf{u})} \right] = \alpha_j - \mathbf{u}' \boldsymbol{\beta}_j, \quad j = 1, 2, \dots, r-1$$

$$\text{GAM: } \ln \left[\frac{P(Y = y_j | \mathbf{x}, \mathbf{u})}{P(Y > y_j | \mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{p_j(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r p_m(\mathbf{x}, \mathbf{u})} \right] = \alpha_j - \mathbf{u}' \boldsymbol{\beta}_j + g_j(\mathbf{x}), \quad j = 1, 2, \dots, r-1$$

Não há nada de novo no ajuste dos GLM binomiais ou multinomiais e mesmo para os GAM binomiais já existe na literatura a maneira de ajustá-los⁽³⁸⁾, porém para ajustar os GAM multinomiais foi preciso desenvolver o processo de análise para o caso em que se tem uma função não-paramétrica bidimensional.

3.1.4 GAM multinomiais

Com base nos GAM apresentados por Hastie e Tibshirani⁽³⁾, Kelsall e Diggle⁽³⁸⁾ propuseram um algoritmo para estimar o risco espacial através da função Kernel Nadaraya-Watson (K_h) de forma semiparamétrica com a adição de covariáveis não espaciais, quando a resposta é binomial. Para um ajuste que incorpore resposta multinomial, tal algoritmo foi aqui modificado conforme se segue:

Seja η_{ij} o preditor linear para o sujeito i na j -ésima comparação, sendo que $i = 1, \dots, n$ e $j = 1, \dots, r$ em que n é o número de indivíduos no estudo e r , o número de categorias da resposta. A função η_{ij} varia para cada tipo de modelo multinomial (as funções de η_{ij} para cada um dos modelos estão apresentadas no Quadro 1). Considerando as notações apresentadas até aqui, o ajuste GAM semiparamétrico para modelos multinomiais, se dá de acordo com o seguinte algoritmo:

- (a) $\hat{g}_j(\mathbf{x}) = 0$ e $\hat{\beta}_j$ é a estimativa de máxima verossimilhança do ajuste do GLM com $\hat{\eta}_{ij} = \mathbf{u}_i' \hat{\beta}_j$ estimado apenas de forma paramétrica.
- (b) Com base no ajuste em (a), calcula-se $\hat{\eta}_{ij} = \mathbf{u}_i' \hat{\beta}_j + g_j(\mathbf{x})$ e \hat{p}_{ij} , conforme apresentado no Quadro 1.
- (c) Deve-se construir, para cada um dos i indivíduos, uma *variável dependente ajustada* z_{ij} e pesos w_{ij} , que, dependendo do modelo, variam conforme se vê no Quadro 1.
- (d) Ajusta-se, através do algoritmo *backfitting* um modelo aditivo ponderado da forma $Z_j = \mathbf{u}'\beta_j + g_j(\mathbf{x}) + \varepsilon_j$, usando regressão Kernel Nadaraya-Watson:
- Considere $s_{ij} = z_{ij} - \mathbf{u}_i' \hat{\beta}_j$ e ajuste um modelo de regressão kernel ponderado de s_{ij} em \mathbf{x}_i com pesos w_{ij} , para uma banda h :
$$\hat{g}_j(\mathbf{x}) = \sum_{i=1}^n w_{ij} K_h(x - x_i) s_{ij} \bigg/ \sum_{i=1}^n w_{ij} K_h(x - x_i)$$
 - Ajustar um modelo de regressão de $z_{ij} - \hat{g}_j(\mathbf{x}_i)$ nos valores \mathbf{u}_i por mínimos quadrados ponderados com pesos w_{ij} para obter novas estimativas de $\hat{\beta}_j$
 - Repetir os passos (i) e (ii) até as estimativas $\hat{\eta}_{ij}$ convergirem
- (e) Repetir os passos (b)-(d) até as estimativas $\hat{\eta}_{ij}$ convergirem.

Como este algoritmo prevê a aplicação em vários modelos multinomiais, e como em cada modelo a forma de $\hat{\eta}_{ij}$ varia, foi necessário, para cada modelo, obter as diferentes fórmulas para \hat{p}_{ij} (probabilidade do indivíduo i ser da categoria j de resposta), z_{ij} e os pesos w_{ij} , sendo que estes dois últimos foram construídos a partir de suas definições em Hastie e Tibshirani⁽³⁾:

$$z_i = \eta_i^0 + (y_i - \mu_i^0) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_0 \quad \text{e} \quad w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_0^2 V_i^0, \text{ sendo } V \text{ a variância de } Y \text{ em } \mu. \text{ A}$$

forma algébrica de obtê-los encontra-se no Apêndice 10.1.

Quadro 1: Funções utilizadas no algoritmo para obtenção dos GAM semiparamétricos com respostas multinomiais.

| | $\hat{\eta}_{ij}$ | \hat{p}_{ij} | z_{ij} | w_{ij} |
|--------------------|---|--|---|---|
| Politômico | $\log \left[\frac{\hat{p}_{ij}}{\hat{p}_{ir}} \right],$ $j = 1, \dots, r-1$ | $\hat{p}_{ir} = \frac{1}{1 + \sum_{m=1}^{r-1} \exp(\hat{\eta}_{im})}, \text{ e}$ $\hat{p}_{ij} = \exp(\hat{\eta}_{ij}) \times \hat{p}_{ir}, j = 1, \dots, r-1$ | $\hat{\eta}_{ij} + \frac{(y_{ij} - \hat{p}_{ij})}{\hat{p}_{ij}},$ $j = 1, \dots, r-1$ | $\frac{\hat{p}_{ij}}{(1 - \hat{p}_{ij})},$ $j = 1, \dots, r-1$ |
| Cumulativo | $\log \left[\frac{\sum_{l=1}^j \hat{p}_{il}}{\sum_{m=j+1}^r \hat{p}_{im}} \right],$ $j = 1, \dots, r-1$ | $\hat{p}_{i1} = \frac{\exp(\hat{\eta}_{i1})}{1 + \exp(\hat{\eta}_{i1})};$ $\hat{p}_{ij} = \frac{\exp(\hat{\eta}_{ij})}{1 + \exp(\hat{\eta}_{ij})} - \sum_{m=1}^{j-1} \hat{p}_{im},$ $j = 2, \dots, r-1 \text{ e}$ $\hat{p}_{ir} = \frac{\sum_{m=1}^{r-1} \hat{p}_{im}}{\exp(\hat{\eta}_{ir-1})}$ | $\hat{\eta}_{ij} + \frac{(y_{ij} - \hat{p}_{ij})}{\sum_{m=1}^j p_{im}},$ $j = 1, \dots, r-1$ | $\frac{\left(\sum_{m=1}^j p_{im} \right)^2}{p_{ij}(1 - p_{ij})},$ $j = 1, \dots, r-1$ |
| Razões Sequenciais | $\log \left[\frac{\hat{p}_{ij}}{\sum_{m=j+1}^r \hat{p}_{im}} \right],$ $j = 1, \dots, r-1$ | $\hat{p}_{i1} = \frac{\exp(\hat{\eta}_{i1})}{1 + \exp(\hat{\eta}_{i1})};$ $\hat{p}_{ij} = \frac{\exp(\hat{\eta}_{ij})}{1 + \exp(\hat{\eta}_{ij})} \times \left(1 - \sum_{m=1}^{j-1} \hat{p}_{im} \right),$ $j = 2, \dots, r-1 \text{ e}$ $\hat{p}_{ir} = \frac{\hat{p}_{ir-1}}{\exp(\hat{\eta}_{ir-1})}$ | $\hat{\eta}_{ij} + \frac{(y_{ij} - \hat{p}_{ij})}{\hat{p}_{ij}},$ $j = 1, \dots, r-1$ | $\frac{\hat{p}_{ij}}{(1 - \hat{p}_{ij})},$ $j = 1, \dots, r-1$ |
| Cat. Adjacentes | $\log \left[\frac{\hat{p}_{ij}}{\hat{p}_{ij+1}} \right],$ $j = 1, \dots, r-1$ | $\hat{p}_{ir} = \frac{1}{1 + \sum_{l=1}^{r-1} \exp \left(\sum_{m=l}^{r-1} \hat{\eta}_{im} \right)}, \text{ e}$ $\hat{p}_{ij} = \exp \left(\sum_{m=j}^{r-1} \hat{\eta}_{im} \right) \times \hat{p}_{ir}, j = 1, \dots, r-1$ | $\hat{\eta}_{ij} + \frac{(y_{ij} - \hat{p}_{ij})}{\hat{p}_{ij}},$ $j = 1, \dots, r-1$ | $\frac{\hat{p}_{ij}}{(1 - \hat{p}_{ij})},$ $j = 1, \dots, r-1$ |

$i = 1, \dots, n$; n é o número de indivíduos no estudo e r é o número de categorias de resposta.

Para a convergência dos $\hat{\eta}_{ij}$ nos passos (d)iii) e (e) adaptou-se o critério apresentado por Hastie e Tibshirani⁽³⁾:

$$\Delta(\hat{\eta}_j^1, \hat{\eta}_j^0) = \frac{\sum_{i=1}^n \|\hat{\eta}_{ij}^1 - \hat{\eta}_{ij}^0\|}{\sum_{i=1}^n \|\hat{\eta}_{ij}^0\|} < \varepsilon, \text{ em que } \varepsilon > 0 \text{ é um valor arbitrário, sendo que o usuário}$$

poderá escolher o valor de ε para considerar a convergência. Nas análises apresentadas nesta tese, o valor considerado foi de 0,0001, sendo que, se este não fosse alcançado, o processo era considerado como tendo convergido na trigésima repetição.

Escolha dos valores de banda

O uso do GAM, em conjunto com as estimativas de kernel, prevê a escolha da banda. Essa escolha pode ser baseada em algum aspecto epidemiológico (alcance de voo do mosquito, por exemplo), visual (o pesquisador pondera sobre qual escolha foi suave o bastante, mas sem perder a variação dos eventos) ou em métodos analíticos como o que se segue.

Optou-se aqui por uma fórmula de validação cruzada (CV) de mínimos quadrados para o passo da regressão não-paramétrica de cada iteração. No passo (d)i) do algoritmo apresentado escolhe-se o valor de h que minimiza:

$$CV(h) = n^{-1} \sum_{i=1}^n w_i \{z_i - \hat{g}_h^{-i}(\mathbf{x}_i)\}^2$$

Onde, $\hat{g}_h^{-i}(\mathbf{x}_i)$ é a estimativa $\hat{g}(\mathbf{x}_i)$ construída com a largura de banda h usando todos os dados, exceto $(\mathbf{x}_i, \mathbf{u}_i)$.

Para o cálculo do $CV(h)$ escolhe-se alguns valores para h e então se analisa qual dos resultados minimizou o valor de CV.

Nas análises realizadas neste trabalho, após a obtenção dos valores de banda através de CV, para fins de comparação e melhor ilustração dos métodos apresentados, optou-se por fixar os valores de banda de todas as análises em $h = 800$.

Análise dos resíduos dos GAM multinomiais

A análise dos resíduos é fundamental em qualquer ajuste de modelo de regressão, pois serve para averiguar se os pressupostos adotados são válidos, possibilitando assim, que se infira sobre os resultados obtidos.

Dentre as análises usuais deve-se considerar a análise dos pressupostos de normalidade e de homocedasticidade dos erros, além de se averiguar a existência de pontos discrepantes. Para tanto, fez-se uso das análises gráficas propostas por Hosmer e Lemeshow⁽⁴⁰⁾. Como nos GAM não se pressupõe normalidade, serão realizadas apenas observações sobre homocedasticidade e pontos discrepantes através de gráficos da dispersão dos resíduos pela probabilidade estimada e da dispersão de resíduos pelo índice de observações no estudo, respectivamente.

Nos casos dos GLM e dos GAM não se pode examinar os resíduos brutos ($\hat{e}_i = y_i - \hat{\mu}_i$, onde Y é a variável resposta e μ é a estimativa referente à resposta). Existem várias sugestões para o cálculo de resíduos que se possa analisar nesses casos, porém aqui se optou pela escolha dos resíduos de Pearson e dos resíduos *deviance*⁽⁴¹⁾.

Resíduos de Pearson

Por definição, este tipo de resíduo corresponde ao resíduo bruto dividido pela raiz da variância do parâmetro estimado: $\hat{e}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{\mu}_i)}}$.

No caso dos modelos logísticos, Y é uma variável indicadora e μ é a probabilidade estimada pelo modelo. Especificadamente no caso dos modelos com resposta multinomial, cada um dos $(r-1)$ logitos estimados possui um vetor de resíduos correspondente e, sendo a variância de $\hat{\mu}$, $n \times \hat{\mu}_j \times (1 - \hat{\mu}_j)$ no j -ésimo logito, os resíduos de Pearson são definidos como:

$$\hat{e}_{ij}^P = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{n \times \hat{\mu}_{ij} \times (1 - \hat{\mu}_{ij})}}, \quad i=1, \dots, n \quad e \quad j=1, \dots, r,$$

onde r é o número de categorias de resposta.

Resíduos Deviance

Para se obter uma análise mais parecida com aquela utilizada em modelos lineares, utilizam-se os resíduos *deviance* em complemento aos resíduos de Pearson por este possuir distribuição simétrica.

O valor de *deviance* é definido no caso multinomial por:

$$D = \sum_i \sum_j d_{ij} \quad (\text{i})$$

Sabe-se que $D = -2 \times \text{Log-Verossimilhança}(\hat{\mu})$ (ii)

e que $\text{Log-Verossimilhança}(\hat{\mu}_{ij}) = \sum_i \sum_j y_{ij} \times \log(\hat{\mu}_{ij})$ (iii).

De (i), (ii) e (iii), obtém-se:

$$d_{ij} = -2 \times y_{ij} \times \log(\hat{\mu}_{ij}).$$

Os resíduos *deviance* são definidos como:

$$\hat{e}^D_{ij} = \text{sinal}(y_{ij} - \hat{\mu}_{ij}) \times \sqrt{d_{ij}}.$$

Como $\hat{\mu}$ é uma probabilidade, e y_{ij} é um indicador, o sinal do resíduo sempre será positivo quando $y_{ij} = 1$ e será negativo, quando $y_{ij} = 0$.

3.2 Análise de variáveis não-espaciais

A matriz de variáveis não-espaciais, denominada neste capítulo por \mathbf{u} , quando inserida em modelos de regressão logísticos, sejam binominais ou multinomiais (seção 3.1.3), permite a estimação de parâmetros β .

Estes parâmetros que podem ser obtidos tanto no ajuste de GLM quanto de GAM, provêm o logaritmo natural da razão de probabilidades de eventos para uma determinada exposição, por isso que ao se extrair a exponencial dos parâmetros estimados nestes modelos se obtém a razão de chances, uma medida epidemiológica.

A RC foi estimada para todas as variáveis não-espaciais dos conjuntos de dados estudados através de modelos binomiais e multinomiais, tanto através do ajuste de GLM quanto de GAM.

O procedimento que se seguiu para a inserção ou exclusão das variáveis em cada modelo foi:

- 1) Realizaram-se ajustes simples de GLM, ou seja, GLM com uma covariável de cada vez;
- 2) Se uma dada covariável obtinha um valor-P inferior a 0,25 no ajuste simples em, pelo menos um dos logitos, ela era incluída no ajuste múltiplo;

3) Após o ajuste do modelo múltiplo, verificava-se as covariáveis com valor-P superior a 0,05, em todos os logitos, e excluía-se, uma a uma, seguindo a ordem de maior a menor valor-P;

4) O modelo múltiplo ficava então com todas as covariáveis que tivessem, em pelo menos um logito, o valor-P inferior a 0,05;

5) A partir deste modelo múltiplo, testou-se interações, uma a uma, com todas as covariáveis que estavam neste modelo;

6) No modelo com interações não se retirou mais nenhuma covariável, mas as interações foram sendo excluídas seguindo o mesmo critério em (3).

7) Tendo o GLM com todas covariáveis e interações significativas, incluía-se as informações do espaço através do ajuste do GAM.

As análises das variáveis espaciais, ou seja, das localizações geográficas dos indivíduos do estudo, serão conduzidas de acordo com os métodos apresentados nas próximas seções (3.3 e 3.4).

3.3 Análise espacial do risco para dados com resposta binomial

Para uma determinada doença e uma delimitada região \mathfrak{R} de interesse (uma cidade, um distrito, um bairro), a função do risco espacial definida por Bithell⁽⁴²⁾ é uma função bivariada descrita da seguinte forma:

$$\theta(\mathbf{x}) = \frac{\lambda(\mathbf{x})}{\pi(\mathbf{x})} \quad (2)$$

em que \mathbf{x} é uma matriz de coordenadas geográficas em \mathfrak{R} , $\lambda(\mathbf{x})$ é a função densidade espacial de ocorrência de casos da doença em \mathfrak{R} (número de ocorrências por unidade de área na localização \mathbf{x}) e $\pi(\mathbf{x})$ é a função densidade espacial da população fonte de casos em \mathfrak{R} .

O risco em (2), para uma dada localização $\mathbf{x} \in \mathfrak{R}$ representa o risco de ocorrência de um caso em \mathbf{x} relativamente ao risco médio de ocorrência em \mathfrak{R} para uma determinada exposição. Se este for conhecido, pode ser usado para se obter o risco absoluto em \mathbf{x} , conforme será visto adiante.

A partir de um conjunto observado de casos e controles, Kelsall e Diggle⁽³⁸⁾ e Diggle⁽⁴³⁾ propuseram um modo interessante de estimar $\theta(\mathbf{x})$, abaixo descrito.

Sejam C e PF os conjuntos de localizações (coordenadas geográficas) de casos e população-fonte em estudo, respectivamente. Assume-se que os elementos de C e PF são realizações de dois processos planares de Poisson, com intensidade $\lambda(\mathbf{x})$ e $\pi(\mathbf{x})$, respectivamente.

Sejam \mathbf{x}_i , $i=1,2,\dots,n_1$, n_1 localizações observadas no conjunto C e \mathbf{x}_i , $i=n_1+1, n_1+2, \dots, n_1+n_2$, n_2 localizações observadas no conjunto PF, em que $n=n_1+n_2$.

Seja y_i um indicador associado ao ponto \mathbf{x}_i , tal que $y_i=1$ se $\mathbf{x}_i \in C$ e $y_i=0$ se $\mathbf{x}_i \in PF$.

Assume-se que os casos e os controles incluídos no estudo são amostras aleatórias de todos os casos ocorridos (C) em \mathfrak{R} e da população-fonte (PF) existente em \mathfrak{R} no período de estudo, nas proporções não conhecidas q_1 e q_2 , respectivamente. Desse modo, condicionalmente aos pontos \mathbf{x}_i , y_i são realizações mutuamente independentes de uma variável aleatória de Bernoulli⁽⁴⁴⁾, onde a probabilidade de qualquer ponto, observado no conjunto de casos e controles, ser um caso, isto é, $p(\mathbf{x})=P(y_i=1|\mathbf{x}_i=\mathbf{x})$, é dada por:

$$p(\mathbf{x}) = \frac{q_1 \lambda(\mathbf{x})}{q_1 \lambda(\mathbf{x}) + q_2 \pi(\mathbf{x})} \quad (3)$$

A partir das equações (2) e (3), com álgebra simples demonstra-se que $p(\mathbf{x})$ está relacionado ao risco espacial $\theta(\mathbf{x})$ conforme a expressão abaixo:

$$\ln[\theta(\mathbf{x})] = \ln\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right] + \ln\left[\frac{q_2}{q_1}\right] = \text{logito}[p(\mathbf{x})] + c = \rho(\mathbf{x}) \quad (4)$$

Desse modo, o logaritmo da função risco espacial, $\rho(\mathbf{x})$, a parte uma constante aditiva c , depende apenas do logito da função $p(\mathbf{x})$, definido na equação (3) e para obter o risco relativo, basta dividir o risco pontual pelo risco médio na região do estudo^(42, 45).

Assim, a tarefa de modelar $\theta(\mathbf{x})$ se reduz efetivamente à modelagem do $\text{logito}[p(\mathbf{x})]$ a partir de um conjunto observado de desfechos Bernoulli, y_i ,

associados às suas localizações espaciais, \mathbf{x}_i . Em essência, este é um problema trivial de regressão logística⁽³⁸⁾, embora aqui, dado o provável comportamento não linear da superfície espacial de risco a ser estimada, o uso de um método não-paramétrico seja mais apropriado do que a usual abordagem paramétrica. Para tanto, Kelsall e Diggle⁽³⁸⁾ propuseram o uso do GAM⁽³⁾ como ferramenta básica e a obtenção de $\rho(\mathbf{x})$ poderia ser feita através da função logito apresentada na seção 3.1.3 para dados binomiais do GAM. O log-risco $\rho(\mathbf{x}, \mathbf{u})$ é assim modelado linearmente em termos de um componente não-espacial paramétrico representando o efeito de covariáveis e não linearmente em termos do componente espacial não paramétrico $g(\mathbf{x})$.

Finalmente, a estimativa da função risco espacial, definida pela equação (2), em \mathfrak{R} é obtida como:

$$\hat{\theta}(\mathbf{x}, \mathbf{u}) = e^{\hat{\rho}(\mathbf{x}, \mathbf{u})}$$

em que $\hat{\rho}(\mathbf{x}, \mathbf{u})$ representa a estimativa obtida no ajuste de $\rho(\mathbf{x}, \mathbf{u})$.

Para dados com resposta multinomial em que se deseje analisar o risco espacial foi desenvolvido um processo de análise semelhante a este e que se encontra na próxima seção.

3.4 Análise espacial do risco para dados com resposta multinomial

Esta seção traz todo o processo para a análise espacial do risco quando se possui um desfecho com mais de duas categorias de interesse. Tal processo considera a possibilidade de adicionar ao estudo do risco espacial a análise de outras informações pertinentes ao fenômeno epidemiológico em questão.

Para conduzir esta nova análise espacial do risco epidemiológico, primeiramente é preciso defini-lo formalmente quando se tem uma variável resposta com distribuição de probabilidade multinomial e indicar qual a melhor metodologia estatística para estimá-lo (seção 3.4.1).

A partir daí, é preciso estabelecer uma conexão entre as estimativas obtidas através dos modelos estatísticos e o risco relativo espacial para estudos com variável resposta multinomial (seção 3.4.2).

O estudo das estimativas dos riscos espaciais só estará completo após a inserção de informações sobre a significância estatística destas estimativas (seção 3.4.3).

Em função de esta medida ser definida na literatura sob o termo *risco espacial*, tal nome será mantido nas ilustrações, independentemente do tipo de estudo epidemiológico conduzido.

3.4.1 Definição do risco espacial multinomial

As definições de risco espacial a seguir se embasam na definição do risco espacial para dados com resposta binomial apresentada na seção 3.3.

Além das notações já utilizadas até esse ponto do capítulo 3, considere q_j , $j=1,2,\dots,r$ proporções desconhecidas de pontos que constituem r realizações parciais de processos independentes de Poisson na região de estudo. Kelsall e Diggle⁽³⁸⁾ argumentam que para muitos conjuntos de dados estas proporções são iguais a um, mas este não será o caso se o estudo for viesado ou se um dos conjuntos de pontos forem controles extraídos aleatoriamente da população sob risco⁽⁴⁶⁾. Mostraremos no item 3.4.2 que, para análises em que o grupo controle é utilizado como referência nas comparações, não é necessário se conhecer os valores q_j explicitamente. Em estudos onde foi realizada uma amostra aleatória da população fonte e somente depois os sujeitos foram classificados em uma das r categorias de resposta, como em estudos transversais, todos os q_j podem ser assumidos como sendo iguais e, neste caso, não se precisa saber exatamente os valores.

Para uma determinada doença e uma delimitada região \mathfrak{R} de interesse, seja \mathbf{x} uma matriz de coordenadas geográficas que mapeia indivíduos com a doença em \mathfrak{R} e \mathbf{u} uma matriz de covariáveis de interesse associadas aos indivíduos estudados. A probabilidade de qualquer indivíduo, observado no conjunto de dados, pertencer a categoria j , é dada por:

$$p_j(\mathbf{x}, \mathbf{u}) = \frac{q_j \lambda_j(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q_m \lambda_m(\mathbf{x}, \mathbf{u})}, j=1, 2, \dots, r \quad (5)$$

sendo $\lambda_j(\mathbf{x})$, a função densidade espacial de ocorrência de eventos do tipo j em

$$\Re \text{ e } \sum_{j=1}^r p_j(\mathbf{x}, \mathbf{u}) = 1.$$

Os riscos espaciais multinomiais a seguir foram definidos a partir dos modelos logísticos cumulativo, de razões sequenciais, de categorias adjacentes e o politômico (apresentados na seção 3.1.3). Isso foi feito em razão da possibilidade de utilizar tais modelos para estimar os riscos espaciais.

Considera-se aqui que a soma de processos de Poisson independentes é também um processo de Poisson com intensidade igual à soma das intensidades de cada processo somado⁽³²⁾.

Análise Cumulativa

Uma forma muito utilizada para análise de dados ordinais, que poderia ser também utilizada na análise espacial é quando se acumulam categorias de resposta, ou seja, em cada análise consideram-se todos os dados do estudo e observa-se o efeito de mudar o nível de comparação estabelecido pelas categorias de resposta. Por exemplo, num estudo de gravidade de uma determinada doença, onde temos o desfecho categorizado em casos *leves*, *moderados*, *graves* e *fatais*, podemos analisar os dados comparando casos *leves* com os demais; depois, casos *leves* e *moderados* em relação aos mais *graves* e *fatais* e, por fim, os casos *fatais* com todos os demais.

Nesta análise, o log-risco é definido da seguinte maneira:

$$\rho_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\sum_{l=1}^j \lambda_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \lambda_m(\mathbf{x}, \mathbf{u})} \right), \quad j = 1, 2, \dots, r-1 \quad (6)$$

Análise de Razões Sequenciais

Pode haver circunstâncias em que haja o interesse em acumularem-se as categorias para a análise em apenas uma direção, ou seja, comparar uma categoria específica com as suas “maiores” ou “menores”. Utilizando nosso

exemplo, faríamos três comparações: casos *fatais* com os outros três menos graves; casos *graves* com os *moderados* e *leves*; e casos *moderados* com *leves*.

Neste caso, o log-risco é definido como:

$$\rho_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \lambda_m(\mathbf{x}, \mathbf{u})} \right), \quad j = 1, 2, \dots, r-1 \quad (7)$$

Análise de Categorias Adjacentes

Muitas vezes pode-se haver a necessidade de não se agregar categorias de resposta, mas compará-las duas a duas. Uma opção conveniente no caso ordinal é comparar categorias mais próximas de resposta. Voltando ao nosso exemplo, este tipo de análise observaria as relações entre os casos *fatais* e *graves*; entre *graves* e *moderados* e entre *moderados* e *leves*.

Neste caso, o log-risco é definido como:

$$\rho_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\lambda_{j+1}(\mathbf{x}, \mathbf{u})} \right), \quad j = 1, 2, \dots, r-1 \quad (8)$$

Análise Nominal e/ou com uma categoria de referência

Quando temos uma resposta nominal, o formato mais indicado de análise é não juntar categorias, mas compará-las duas a duas. Por exemplo, um estudo sobre doenças respiratórias, em que há como resposta as categorias *pneumonias*, *bronquites* e *infecções*. A melhor maneira de fazer tal análise é eleger uma delas como a de referência e comparar todas as outras com esta. Quando há uma categoria de controle, esta é a melhor opção para ser a referência.

Seja r a categoria de referência para as demais comparações, o log-risco é definido da seguinte maneira:

$$\rho_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\lambda_r(\mathbf{x}, \mathbf{u})} \right), \quad j = 1, 2, \dots, r-1 \quad (9)$$

O Quadro 2 traz um resumo dos riscos espaciais multinomiais apresentados relacionando-os aos respectivos modelos logísticos que devem ser utilizados para obter suas estimativas e também uma breve descrição do uso de cada uma das análises citadas acima.

Quadro 2: Análises espaciais multinomiais com respectivas funções de risco espacial e modelo logístico.

| Análise Multinomial | Modelo Logístico | Risco Espacial* | Uso |
|-----------------------|-----------------------|--|---|
| Cumulativa | Cumulativo | $p_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\sum_{l=1}^j \lambda_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \lambda_m(\mathbf{x}, \mathbf{u})} \right)$ | Quando se tem variável resposta ordinal proveniente de estudo transversal. <i>Vantagem:</i> maior consistência da informação. <i>Desvantagem:</i> não se tem informação de risco para categorias específicas da variável resposta, pois acumula as categorias na análise. |
| Razões Sequenciais | Razões Sequenciais | $p_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \lambda_m(\mathbf{x}, \mathbf{u})} \right)$ | Quando se tem variável resposta ordinal proveniente de estudo transversal. <i>Vantagem:</i> equilibra a especificidade em que se têm riscos para todas as categorias da variável resposta com a consistência da informação. <i>Desvantagem:</i> não é tão consistente como as informações da análise cumulativa, nem tão específico quando os riscos da análise de categorias adjacentes. |
| Categorias Adjacentes | Categorias Adjacentes | $p_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\lambda_{j+1}(\mathbf{x}, \mathbf{u})} \right)$ | Quando se tem variável resposta ordinal proveniente de estudo transversal. <i>Vantagem:</i> provê medidas de risco e/ou razão de chances para cada categoria da variável resposta separadamente. <i>Desvantagem:</i> para amostras pequenas ou estudos em que se têm poucos indivíduos em alguma categoria de estudo, não é uma análise muito consistente. |
| Nominal | Politômico | $p_j(\mathbf{x}, \mathbf{u}) = \ln \left(\frac{\lambda_j(\mathbf{x}, \mathbf{u})}{\lambda_r(\mathbf{x}, \mathbf{u})} \right)$ | Pode-se utiliza-la sempre que se tem uma variável resposta multinomial, porém é mais indicada quando se tem variável resposta nominal e/ou quando o estudo epidemiológico exige que se fixe uma categoria como referência para as análises. <i>Vantagem:</i> obtém todas as medidas de risco comparando com uma só categoria de referência, o que facilita a discussão. <i>Desvantagem:</i> como na análise de categorias adjacentes, nem sempre proverá medidas muito consistentes e para casos em que a variável resposta é ordinal, as demais análises provêm medidas de risco mais intuitivas de se discutir. |

* Todos os risco espaciais definidos para $j=1,2,\dots,r-1$, sendo r o número de categorias da variável resposta.

O processo de estimação do risco espacial apresentado na seção 3.3 é descrito na literatura para o caso de dados com resposta binomial. Para estender esse processo para ocasiões em que a resposta é multinomial, ajustam-se os

GAM propostos na seção 3.1.4 e, a partir desses ajustes, calcula-se o risco relativo espacial multinomial como se apresenta na próxima seção.

3.4.2 Risco relativo espacial multinomial

Em casos em que se parte de uma amostra aleatória de indivíduos cujas respostas são classificadas em categorias, pode-se pressupor que todas as frações amostrais são iguais à fração amostral do estudo ($q_j = q$) e, em situações em que se estudam todos os casos incidentes ou prevalentes, $q=1$. Nestes casos tem-se:

$$\hat{p}_j(\mathbf{x}, \mathbf{u}) = \frac{q\hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q\hat{\lambda}_m(\mathbf{x}, \mathbf{u})} = \frac{\hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r \hat{\lambda}_m(\mathbf{x}, \mathbf{u})}, j=1, 2, \dots, r$$

Vê-se que, nesses casos especiais, a estimação dos log-riscos definidos nas equações (6)–(9) são os próprios logitos estimados sob os GAM.

Para se extrair estimativas de risco da categoria j de resposta, para um determinado ponto i definido dentro de \mathfrak{R} , a partir das estimativas dos GAM, basta obter os valores exponenciados dos log-riscos estimados no ponto i , ou seja:

$$\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u}) = \exp(\hat{\rho}_{ij}(\mathbf{x}, \mathbf{u})), \quad j=1, \dots, r-1; i=1, \dots, n$$

Para obter o risco relativo, divide-se o risco estimado em cada ponto pela média dos riscos para a região:

$$\hat{\Theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\frac{\sum_{j=1}^n \hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{n}}, \quad j=1, \dots, r-1; i=1, \dots, n$$

Há duas análises em que, mesmo não pressupondo igualdade entre as frações amostrais, pode-se obter o risco espacial diretamente da estimativa do logito sob os GAM: a análise de categorias adjacentes e a análise nominal.

Análises de Categorias Adjacentes

Nesta análise, temos o risco estimado, a menos uma constante multiplicativa, como:

$$c_{j1} \times \hat{\theta}_j(\mathbf{x}, \mathbf{u}) = \exp \left(\ln \left[\frac{\hat{p}_j(\mathbf{x}, \mathbf{u})}{\hat{p}_{j+1}(\mathbf{x}, \mathbf{u})} \right] \right) = \frac{\left[\frac{q_j \hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right]}{\left[\frac{q_{j+1} \hat{\lambda}_{j+1}(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right]} = \frac{q_j}{q_{j+1}} \times \frac{\hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{j+1}(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1,$$

onde $c_{j1} = q_j / q_{j+1}$.

Daí, o risco relativo espacial é obtido da seguinte maneira:

$$\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\frac{c_{j1} \times \hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{l=1}^n c_{j1} \times \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}}{\frac{c_{j1} \times \sum_{l=1}^n \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}{n}} = \frac{\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{l=1}^n \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1; i=1, \dots, n$$

Análise Nominal

O logito estimado para este tipo de análise também permite obter o risco estimado, a menos uma constante multiplicativa, da seguinte maneira:

$$c_{j2} \times \hat{\theta}_j(\mathbf{x}, \mathbf{u}) = \exp \left(\ln \left[\frac{\hat{p}_j(\mathbf{x}, \mathbf{u})}{\hat{p}_r(\mathbf{x}, \mathbf{u})} \right] \right) = \frac{\left[\frac{q_j \hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right]}{\left[\frac{q_r \hat{\lambda}_r(\mathbf{x}, \mathbf{u})}{\sum_{m=1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right]} = \frac{q_j}{q_r} \times \frac{\hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_r(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1,$$

sendo $c_{j2} = q_j / q_r$.

O risco relativo espacial é obtido de forma análoga à análise de categorias adjacentes:

$$\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\frac{c_{j2} \times \hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{l=1}^n c_{j2} \times \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}}{\frac{c_{j2} \times \sum_{l=1}^n \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}{n}} = \frac{\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{l=1}^n \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1; i=1, \dots, n$$

Porém, nas outras duas análises sugeridas é preciso conhecer as frações amostrais para se obter os riscos relativos espaciais da maneira que se segue.

Análises Cumulativas

O logito estimado pelo GAM tem aqui a seguinte forma:

$$\ln \left[\frac{\sum_{l=1}^j \hat{p}_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \hat{p}_m(\mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{\sum_{l=1}^j q_l \hat{\lambda}_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right], \quad j=1, 2, \dots, r-1$$

Em razão do formato cumulativo, é necessária a resolução de um sistema linear de equações para conseguir separar as frações amostrais (q_j) das funções de densidade espacial ($\lambda(\mathbf{x}, \mathbf{u})$):

$$\begin{pmatrix} \frac{q_2}{q_1} \hat{\phi}_{i1}(\mathbf{x}, \mathbf{u}) & \frac{q_3}{q_1} \hat{\phi}_{i1}(\mathbf{x}, \mathbf{u}) & \dots & \frac{q_r}{q_1} \hat{\phi}_{i1}(\mathbf{x}, \mathbf{u}) \\ -\frac{q_2}{q_1} & \frac{q_3}{q_1} \hat{\phi}_{i2}(\mathbf{x}, \mathbf{u}) & \dots & \frac{q_r}{q_1} \hat{\phi}_{i2}(\mathbf{x}, \mathbf{u}) \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{q_2}{q_1} & -\frac{q_3}{q_1} & \dots & \frac{q_r}{q_1} \hat{\phi}_{ir-1}(\mathbf{x}, \mathbf{u}) \end{pmatrix} \begin{pmatrix} \hat{\lambda}_{i2}(\mathbf{x}, \mathbf{u}) \\ \hat{\lambda}_{i1}(\mathbf{x}, \mathbf{u}) \\ \hat{\lambda}_{i3}(\mathbf{x}, \mathbf{u}) \\ \vdots \\ \hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u}) \\ \hat{\lambda}_{i1}(\mathbf{x}, \mathbf{u}) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad i=1, \dots, n \quad (10)$$

Seja $\hat{\phi}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\sum_{l=1}^j q_l \hat{\lambda}_{il}(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r q_m \hat{\lambda}_{im}(\mathbf{x}, \mathbf{u})}$, a exponencial do logito estimado pelo GAM para

o ponto i .

Como o determinante da matriz do sistema linear é $\frac{q_2}{q_1} \dots \frac{q_r}{q_1} \hat{\phi}_{i1}(\mathbf{x}, \mathbf{u}) (\hat{\phi}_{i2}(\mathbf{x}, \mathbf{u}) + 1) \dots (\hat{\phi}_{ir-1}(\mathbf{x}, \mathbf{u}) + 1)$, e sendo cada q_i e $\phi_{ij}(\mathbf{x}, \mathbf{u})$ positivos, o sistema sempre terá uma única solução.

O desenvolvimento algébrico para a obtenção de tal sistema linear encontra-se no anexo 10.2.

A resolução dos sistemas lineares expressos na equação (10) provê as razões das funções de densidade espacial, para cada ponto i , suficientes para obter os riscos espaciais da seguinte maneira:

$$\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{1 + \sum_{l=2}^j \frac{\hat{\lambda}_l(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{\sum_{m=j+1}^r \frac{\hat{\lambda}_m(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} = \frac{\sum_{l=1}^j \frac{\hat{\lambda}_l(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{\sum_{m=j+1}^r \frac{\hat{\lambda}_m(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} = \frac{\sum_{l=1}^j \hat{\lambda}_l(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \hat{\lambda}_m(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1; i=1, \dots, n.$$

Análise de Razões Sequenciais

Assim como a análise cumulativa, esta também necessita de um cálculo diferente para extrair o risco a partir do logito estimado pelo GAM:

$$\ln \left[\frac{\hat{p}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \hat{p}_m(\mathbf{x}, \mathbf{u})} \right] = \ln \left[\frac{q_j \hat{\lambda}_j(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r q_m \hat{\lambda}_m(\mathbf{x}, \mathbf{u})} \right], \quad j=1, 2, \dots, r-1$$

Conhecendo-se as razões amostrais, obtém-se a exponencial do $(r-1)$ -ésimo logito estimado:

$$\hat{\phi}_{ir-1}(\mathbf{x}, \mathbf{u}) = \frac{q_{r-1} \hat{\lambda}_{ir-1}(\mathbf{x}, \mathbf{u})}{q_r \hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})} \quad \text{e} \quad \hat{\theta}_{ir-1}(\mathbf{x}, \mathbf{u}) = \frac{\hat{\phi}_{ir-1}(\mathbf{x}, \mathbf{u}) q_r}{q_{r-1}} = \frac{\hat{\lambda}_{ir-1}(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})}.$$

Com esta medida obtém-se:

$$\frac{\hat{\lambda}_{ij}(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})} = \frac{\hat{\phi}_{ij}(\mathbf{x}, \mathbf{u}) \left(\sum_{l=j+1}^{r-1} q_l \frac{\hat{\lambda}_{il}(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})} + q_r \right)}{q_j}, \quad j=r-2, \dots, 1 \quad \text{e}$$

$$\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\frac{\hat{\lambda}_{ij}(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})}}{\sum_{m=j+1}^r \frac{\hat{\lambda}_{im}(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_{ir}(\mathbf{x}, \mathbf{u})}} = \frac{\hat{\lambda}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{m=j+1}^r \hat{\lambda}_{im}(\mathbf{x}, \mathbf{u})}, \quad j=1, 2, \dots, r-1.$$

Para ambas análises,

$$\hat{\Theta}_{ij}(\mathbf{x}, \mathbf{u}) = \frac{\hat{\theta}_{ij}(\mathbf{x}, \mathbf{u})}{\sum_{l=1}^n \hat{\theta}_{lj}(\mathbf{x}, \mathbf{u})}, \quad j=1, \dots, r-1; i=1, \dots, n$$

3.4.3 Intervalo de confiança para o risco relativo espacial

A significância dos componentes paramétricos é obtida de modo usual, como nos GLM, com base na magnitude da razão entre estimativas e seus respectivos erros padrão. Entretanto, para obter a significância das estimativas espaciais não foram ainda descritos métodos analíticos. Por isso aqui é apresentado um método adaptado para esta situação.

Kelsall e Diggle⁽³⁸⁾ propõem um procedimento de Monte Carlo para dados binomiais, aqui adaptado para a análise multinomial e descrito abaixo:

a) O modelo completo é ajustado aos dados via GAM. Desse modo, predições de respostas são obtidas para cada ponto de uma grade previamente definida (300 x 300, por exemplo) cobrindo \mathfrak{R} e, destas, são obtidos os riscos espaciais definidos na seção 3.4.1.

b) O modelo reduzido que omite o componente espacial $g(\mathbf{x})$ (GLM), já ajustado por meio da função *vglm* do pacote *VGAM*, é agora utilizado para fazer predições da probabilidade de cada indivíduo do estudo ser classificado em uma determinada categoria de resposta de acordo com o componente não-espacial ($\mathbf{u}'\boldsymbol{\beta}$).

c) Para cada um dos sujeitos estudados é determinado um novo *status* de resposta dentre as categorias estudadas de modo aleatório usando-se as probabilidades preditas no passo b) acima.

d) O modelo completo é ajustado aos dados modificados no passo c) via GAM, e assim novos riscos são obtidos para cada ponto da mesma grade previamente definida no passo a).

e) Os passos c) e d) são repetidos um grande número de vezes (por exemplo, 500) produzindo simulações de superfícies resposta sob a hipótese de homogeneidade espacial do risco.

f) Com base nas estimativas obtidas em e), obtêm-se, para cada ponto da grade, valores de corte para o intervalo de confiança de 95%, a partir de uma distribuição Normal.

g) Finalmente, um mapa é produzido com o risco relativo espacial estimado para toda a grade definida inicialmente e, sobre ela, são delineadas as

informações que indicam a significância (quais estimativas estão dentro ou fora da região entre os valores de corte).

3.5 Aspectos computacionais

Todas as análises de modelos estatísticos foram realizadas com o pacote computacional R⁽⁴⁷⁾, na versão 2.10.

Os GLM foram ajustados através da função *glm*, para as análises de modelos com respostas binomiais; e através da função *vglm* do pacote *VGAM*, para as análises de modelos com respostas multinomiais.

Os GAM binomiais e multinomiais foram ajustados através de funções desenvolvidas neste doutorado junto com os pesquisadores Luciana B. Nucci e Celso Stephan. Tanto o código das análises quanto o pacote de funções para estes ajustes no R encontram-se disponíveis na internet através da página do laboratório epiGeo⁽⁴⁸⁾.

Todos os mapas foram construídos no ArcMap, versão 9.2. Os mapas de risco, apesar de serem construídos no ArcMap se baseiam nas estimativas obtidas pelo R.

4. Estudo caso-controle com resposta multinomial

O entendimento de relações causais que poderiam explicar o surgimento de casos de uma doença pode ser feito através da execução de estudos caso-controle. Neles, a distribuição de exposições associadas aos doentes analisados é comparada àquela existente na população onde habitam esses doentes⁽⁴⁹⁾.

Este tipo de investigação propicia uma maneira eficiente de se obter estimativas populacionais de risco relativo e fração atribuível. A razão desta eficiência advém da capacidade que o estudo caso-controle tem, a um custo relativamente baixo, de estimar a distribuição populacional de exposições a partir do seu grupo controle. Esta propriedade dispensa o consumo de tempo, bem como recursos financeiros e administrativos, na enumeração e seguimento de um grande conjunto de indivíduos, classificados de acordo com níveis de exposição⁽⁵⁰⁾.

A análise de dados de estudos caso-controle pressupõe a representação da população fonte de casos por meio do grupo controle. As medidas epidemiológicas de risco e associação resultantes, tais como razão de proporção de incidência, razão de taxa de incidência e fração atribuível sempre comparam os casos com o grupo de controles⁽⁵⁰⁾. Entretanto, quando se tem mais de um grupo de casos, a pressuposição de que os controles representam a população continua a mesma, devendo esse grupo ser a categoria de referência em todas as comparações realizadas na análise. Em caso contrário, as estimativas obtidas com o ajuste do modelo não podem ser utilizadas para a obtenção das medidas epidemiológicas citadas.

Muito embora seja frequente categorizarmos casos de acordo com uma escala ordinal (por exemplo, hipertenso leve, moderado e grave), a melhor escolha para a modelagem da relação entre exposições e ocorrência de casos assim classificados é através do modelo politômico (seção 3.1.3), tendo os controles como categoria de base⁽³⁶⁾. Epidemiologicamente, esta é a única opção, dentre as possibilidades de análise utilizando a distribuição multinomial, que mantém a lógica do estudo caso-controle, onde a distribuição de exposições dos casos incidentes (agora, classificados em diferentes grupos) é comparada à distribuição

de exposições na população que gerou estes casos, estimada por meio do grupo controle.

Isso também ocorre em outros estudos epidemiológicos em que, por definição, há um grupo para referência das análises, como em estudos de intervenção, onde se tem o grupo de placebo ou do tratamento usual, e em estudos do tipo coorte, onde, mesmo que haja vários níveis de exposição, a comparação deve ser feita com o grupo de não-expostos.

Para ilustrar o uso de análise espacial multinomial em dados de estudos caso-controle, analisaram-se os dados a seguir.

4.1 Acidentes de trabalho em Piracicaba, SP.

Utilizaram-se dados provenientes de um estudo caso-controle de base populacional com 779 trabalhadores precarizados¹ que sofreram acidente do trabalho na cidade de Piracicaba, sudeste do Brasil, em 2007, e que foram atendidos em um dos oito pronto-socorros desta cidade. Como controles, foram estudados 1544 trabalhadores precarizados que não sofreram acidente de trabalho nos três anos anteriores e que trabalhassem na região de estudo⁽⁵¹⁾.

O estudo destes dados tem por objetivo entender a distribuição, a caracterização e os determinantes dos acidentes de trabalho para esta classe específica de trabalhadores, isso porque o acesso a informações destes é extremamente restrita dada a subnotificação de acidentes entre os trabalhadores formais e a não existência de notificação de acidentes entre os informais⁽⁵²⁾.

Neste trabalho, para a análise espacial, foram utilizadas as seguintes variáveis:

- coordenadas geográficas do local de ocorrência dos acidentes, para casos e;
- coordenadas geográficas do local de trabalho, para controles;

A variável resposta possui as seguintes classes: casos graves ($n_1 = 225$), casos leves ($n_2 = 554$) e controles ($n_3 = 1544$). A gravidade do caso foi estipulada com base em diagnóstico médico do momento do atendimento. Para a análise

¹ O roteiro para a identificação do que foi considerado trabalhador precarizado está em anexo (Capítulo 9).

binomial utilizou-se uma variável resposta classificada apenas em casos e controles. A Tabela 1 traz a descrição das variáveis não-espaciais consideradas no estudo.

Tabela 1: Médias e desvios padrão das variáveis contínuas e porcentagens de indivíduos em cada classe por categorias de resposta. Acidentes de trabalho em Piracicaba, 2007.

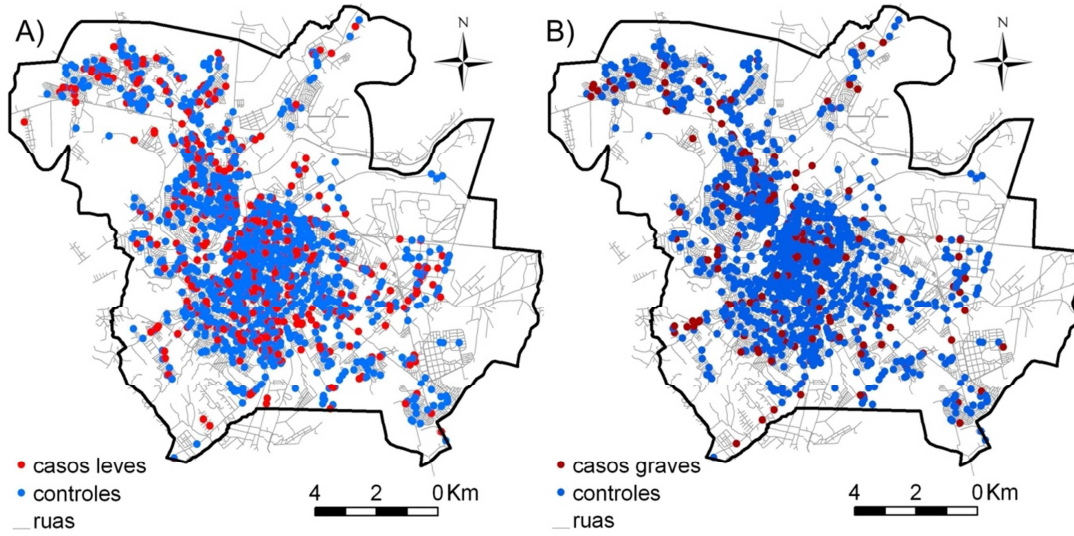
| Variáveis | | Casos Graves (n=225) | Casos Leves (n=554) | Controles (n=1544) | Total (N=2323) |
|-------------------------------|-------|-------------------------|------------------------|-----------------------|-------------------|
| Idade (em anos) | | 35,41 ± 12,68 | 33,11 ± 11,40 | 36,61 ± 11,26 | 35,66 ± 11,53 |
| Escolaridade (em anos) | | 7,45 ± 3,44 | 7,50 ± 3,41 | 7,86 ± 3,58 | 7,73 ± 3,53 |
| Sexo | Fem. | 15,11% | 20,04% | 36,33% | 30,39% |
| | Masc. | 84,89% | 79,96% | 63,67% | 69,61% |
| Possui carteira de trabalho? | Não | 57,33% | 49,10% | 66,19% | 61,26% |
| | Sim | 42,67% | 50,90% | 33,81% | 38,74% |
| Trabalha na rua? | Não | 59,56% | 61,55% | 63,67% | 62,76% |
| | Sim | 40,44% | 38,45% | 36,33% | 37,24% |
| Recebe férias ou 13º salário? | Não | 55,11% | 40,25% | 58,48% | 53,81% |
| | Sim | 44,89% | 59,75% | 41,52% | 46,19% |
| O trabalho é doméstico? | Não | 95,11% | 89,89% | 83,16% | 85,92% |
| | Sim | 4,89% | 10,11% | 16,84% | 14,08% |
| O trabalho é terceirizado? | Não | 77,78% | 71,30% | 82,12% | 79,12% |
| | Sim | 22,22% | 28,70% | 17,88% | 20,88% |

Na Figura 1, tem-se a distribuição espacial de pontos dos indivíduos deste estudo.

Apesar de os casos serem classificados seguindo a ordem da gravidade do acidente, por se tratar de um estudo caso-controle, para sua análise multinomial foi utilizado apenas o modelo politômico, como já justificado na introdução deste capítulo.

O ajuste dos modelos para a análise destes dados seguiu o procedimento descrito na seção 3.2 e os resultados obtidos estão a seguir.

Figura 1: Distribuição espacial dos locais de acidente e trabalho em Piracicaba, 2007. A) Pontos dos casos leves e dos controles. B) Pontos dos casos graves e controles.



4.2 Resultados dos acidentes de trabalho

O modelo binomial múltiplo final permaneceu com as variáveis: escolaridade, idade e sexo do trabalhador e também se este possui carteira de trabalho e se trabalha na rua. Foram significativas as interações entre idade e possuir carteira de trabalho, entre sexo e possuir carteira de trabalho e entre sexo e trabalhar na rua. As funções logito que expressam essas relações são:

$$\text{GLM: } \ln \left[\frac{p_{\text{caso}}(\mathbf{u})}{p_{\text{controle}}(\mathbf{u})} \right] = \alpha + \beta_1 \text{Esc} + \beta_2 \text{Idad} + \beta_3 \text{Sex} + \beta_4 \text{CLT} + \beta_5 \text{Rua} + \\ \beta_6 \text{Idad} \times \text{CLT} + \beta_7 \text{Sex} \times \text{CLT} + \beta_8 \text{Sex} \times \text{Rua}$$

$$\text{GAM: } \ln \left[\frac{p_{\text{caso}}(\mathbf{x}, \mathbf{u})}{p_{\text{controle}}(\mathbf{x}, \mathbf{u})} \right] = \alpha + \beta_1 \text{Esc} + \beta_2 \text{Idad} + \beta_3 \text{Sex} + \beta_4 \text{CLT} + \beta_5 \text{Rua} + \\ \beta_6 \text{Idad} \times \text{CLT} + \beta_7 \text{Sex} \times \text{CLT} + \beta_8 \text{Sex} \times \text{Rua} + g(\mathbf{x})$$

Quanto ao modelo multinomial, além das variáveis significativas ao modelo binomial ainda acresceu-se a informação sobre o trabalho ser doméstico e o indivíduo possuir direitos como férias e recebimento do décimo terceiro salário. As interações significativas foram entre idade e sexo, sexo e possuir carteira de

trabalho, ter direitos trabalhistas e receber carteira de trabalho e entre trabalhar na rua e sexo. As funções logito que expressam essas relações são:

$$\begin{aligned}
 \text{GLM: } \ln \left[\frac{p_{\text{caso grave}}(\mathbf{u})}{p_{\text{controle}}(\mathbf{u})} \right] &= \alpha_1 + \beta_1 \text{Esc} + \beta_2 \text{Idad} + \beta_3 \text{Sex} + \beta_4 \text{CLT} + \beta_5 \text{Rua} + \beta_6 \text{Domest} + \beta_7 \text{DirTrab} + \\
 &\quad \beta_8 \text{Idad} \times \text{Sex} + \beta_9 \text{Sex} \times \text{CLT} + \beta_{10} \text{Sex} \times \text{Rua} + \beta_{11} \text{CLT} \times \text{DirTrab} \\
 \ln \left[\frac{p_{\text{caso leve}}(\mathbf{u})}{p_{\text{controle}}(\mathbf{u})} \right] &= \alpha_2 + \beta_{12} \text{Esc} + \beta_{13} \text{Idad} + \beta_{14} \text{Sex} + \beta_{15} \text{CLT} + \beta_{16} \text{Rua} + \beta_{17} \text{Domest} + \beta_{18} \text{DirTrab} + \\
 &\quad \beta_{19} \text{Idad} \times \text{Sex} + \beta_{20} \text{Sex} \times \text{CLT} + \beta_{21} \text{Sex} \times \text{Rua} + \beta_{22} \text{CLT} \times \text{DirTrab} \\
 \text{GAM: } \ln \left[\frac{p_{\text{caso grave}}(\mathbf{x}, \mathbf{u})}{p_{\text{controle}}(\mathbf{x}, \mathbf{u})} \right] &= \alpha_1 + \beta_1 \text{Esc} + \beta_2 \text{Idad} + \beta_3 \text{Sex} + \beta_4 \text{CLT} + \beta_5 \text{Rua} + \beta_6 \text{Domest} + \beta_7 \text{DirTrab} + \\
 &\quad \beta_8 \text{Idad} \times \text{Sex} + \beta_9 \text{Sex} \times \text{CLT} + \beta_{10} \text{Sex} \times \text{Rua} + \beta_{11} \text{CLT} \times \text{DirTrab} + g(\mathbf{x}) \\
 \ln \left[\frac{p_{\text{caso leve}}(\mathbf{x}, \mathbf{u})}{p_{\text{controle}}(\mathbf{x}, \mathbf{u})} \right] &= \alpha_2 + \beta_{12} \text{Esc} + \beta_{13} \text{Idad} + \beta_{14} \text{Sex} + \beta_{15} \text{CLT} + \beta_{16} \text{Rua} + \beta_{17} \text{Domest} + \beta_{18} \text{DirTrab} + \\
 &\quad \beta_{19} \text{Idad} \times \text{Sex} + \beta_{20} \text{Sex} \times \text{CLT} + \beta_{21} \text{Sex} \times \text{Rua} + \beta_{22} \text{CLT} \times \text{DirTrab} + g(\mathbf{x})
 \end{aligned}$$

No caso dos modelos binomiais, uma análise de *deviance* que compara os modelos GAM e GLM finais nos dá um valor de 54,9 com 1 grau de liberdade, mostrando que a inclusão do espaço como variável explicativa é altamente significativa (valor-P<0,0001). No caso dos modelos multinomiais, a *deviance* é de 87 com 2 graus de liberdade, também mostrando a significância da inclusão do espaço no estudo (valor-P<0,0001).

Quando se tem interações em um modelo logístico, a maneira de obter estimativas de razões de chance muda um pouco: não é simplesmente extrair a exponencial dos parâmetros. Nestes casos não se tem um valor de RC bruto para uma variável que interage com outra, mas sim um RC para cada uma das classes da variável de interação. Como os modelos indicaram a existência de diversas interações escolheram-se as que pareciam mais coerentes para o cálculo das razões de chances. As informações para os componentes paramétricos dos modelos finais estão nas Tabelas 2 e 3 para os modelos binomiais e multinomiais, respectivamente.

Tabela 2: Razões de chances (RC) obtidas através do ajuste do modelo binomial dos dados de acidente de trabalho em Piracicaba, 2007 (N=2323).

| | | GLM | | | GAM | | |
|----------------------------|----------|-------|--------|--------|-------|--------|--------|
| | | RC | LI(RC) | LS(RC) | RC | LI(RC) | LS(RC) |
| Escolaridade (Anos) | | 0,935 | 0,909 | 0,962 | 0,939 | 0,923 | 0,954 |
| Idade (Anos) | CT (Sim) | 0,962 | 0,948 | 0,975 | 0,963 | 0,955 | 0,971 |
| | CT (Não) | 0,982 | 0,965 | 0,999 | 0,979 | 0,969 | 0,990 |
| Carteira de Trabalho (Sim) | Homens | 3,440 | 2,212 | 5,350 | 2,959 | 2,151 | 4,070 |
| | Mulheres | 6,571 | 3,224 | 13,391 | 6,305 | 3,985 | 9,976 |
| Sexo (Masculino) | TR (sim) | 1,099 | 0,654 | 1,846 | 1,022 | 0,726 | 1,438 |
| | TR (não) | 2,166 | 1,527 | 3,072 | 1,960 | 1,580 | 2,431 |

CT: possui carteira de trabalho. TR: trabalho na rua.

RC: Razão de Chances; LI: 95% Limite Inferior; LS: 95% Limite Superior.

GLM: modelo linear generalizado; GAM: modelo aditivo generalizado.

Tabela 3: Razões de chances (RC) obtidas através do ajuste do modelo multinomial dos dados de acidente de trabalho em Piracicaba, 2007 (N=2323).

| | | Logito 1: acidentes graves versus controles | | | | | |
|----------------------------|-----------|---|--------|--------|-------|--------|--------|
| | | GLM | | | GAM | | |
| | | RC | LI(RC) | LS(RC) | RC | LI(RC) | LS(RC) |
| Escolaridade (Anos) | | 0,949 | 0,908 | 0,993 | 0,953 | 0,916 | 0,991 |
| Trabalho doméstico (Não) | | 2,328 | 1,123 | 4,826 | 2,154 | 1,076 | 4,313 |
| Idade (Anos) | Homens | 0,993 | 0,956 | 1,032 | 0,993 | 0,958 | 1,029 |
| | Mulheres | 0,950 | 0,917 | 0,985 | 0,950 | 0,918 | 0,983 |
| Sexo (Masculino) | CT (Não) | 0,788 | 0,357 | 1,741 | 0,794 | 0,381 | 1,656 |
| | CT (Sim) | 0,352 | 0,088 | 1,410 | 0,337 | 0,095 | 1,191 |
| Carteira de Trabalho (Sim) | DIR (Não) | 3,611 | 0,317 | 41,187 | 3,459 | 0,449 | 26,654 |
| | DIR (Sim) | 7,151 | 2,441 | 20,954 | 7,178 | 2,575 | 20,010 |
| Trabalha na rua (Sim) | Homens | 0,851 | 0,311 | 2,324 | 0,921 | 0,363 | 2,338 |
| | Mulheres | 0,756 | 0,292 | 1,958 | 0,753 | 0,310 | 1,832 |

| | | Logito 2: acidentes leves versus controles | | | | | |
|----------------------------|-----------|--|--------|--------|--------|--------|--------|
| | | GLM | | | GAM | | |
| | | RC | LI(RC) | LS(RC) | RC | LI(RC) | LS(RC) |
| Escolaridade (Anos) | | 0,925 | 0,895 | 0,955 | 0,926 | 0,906 | 0,946 |
| Trabalho doméstico (Não) | | 1,136 | 0,754 | 1,711 | 1,092 | 0,803 | 1,484 |
| Idade (Anos) | Homens | 0,972 | 0,950 | 0,994 | 0,970 | 0,952 | 0,987 |
| | Mulheres | 0,967 | 0,947 | 0,987 | 0,966 | 0,950 | 0,982 |
| Sexo (Masculino) | CT (Não) | 3,127 | 1,908 | 5,123 | 3,219 | 2,195 | 4,721 |
| | CT (Sim) | 1,825 | 0,753 | 4,419 | 1,671 | 0,866 | 3,224 |
| Carteira de Trabalho (Sim) | DIR (Não) | 10,363 | 2,415 | 44,466 | 10,241 | 5,691 | 18,428 |
| | DIR (Sim) | 1,856 | 1,105 | 3,118 | 1,999 | 1,346 | 2,967 |
| Trabalha na rua (Sim) | Homens | 0,671 | 0,369 | 1,219 | 0,738 | 0,477 | 1,142 |
| | Mulheres | 1,641 | 0,950 | 2,836 | 1,666 | 1,109 | 2,504 |

CT: possui carteira de trabalho; DIR: possui direitos trabalhistas como férias e décimo terceiro salário.

RC: Razão de Chances; LI: 95% Limite Inferior; LS: 95% Limite Superior.

GLM: modelo linear generalizado; GAM: modelo aditivo generalizado.

Apenas pelo tamanho das tabelas pode-se observar a quantidade de informações que a análise multinomial (Tabela 3) trouxe para a compreensão do agravo estudado em relação à análise binomial (Tabela 2).

A idade e a escolaridade são fatores associados à maior proteção, porém a idade não se mostrou significativa para homens que sofreram acidentes graves e, para casos leves, indica maior proteção para as mulheres.

Quanto ao estudo do sexo dos trabalhadores, observa-se na análise binomial que este é um fator associado à maior risco para homens que não trabalham na rua. Para casos graves, não foi fator significativo, mas, para casos leves, ser do sexo masculino está associado à maior risco se o trabalhador não possuir a carteira de trabalho.

O fato de o trabalhador possuir carteira de trabalho está associado à maior risco. No modelo binomial, o risco de se possuir carteira de trabalho é o dobro para mulheres do que para homens. No modelo multinomial, vemos que este risco aumenta para os casos graves quando combinado ao fato de se ter direitos como férias e recebimento do décimo terceiro salário, mas para casos leves é menor do que se não se ter estes mesmos direitos.

Trabalhar na rua está associado à maior risco para mulheres que sofreram acidentes leves e o trabalho doméstico é fator de risco apenas para casos graves.

Foram ainda feitos gráficos para a análise dos resíduos dos ajustes GAM multinomiais, os quais se encontram no Apêndice 10.3, e mapas das análises espaciais, que estão na seção a seguir.

Análises espaciais

As distribuições espaciais dos RR estimados podem ser visualizadas na Figura 2.

Observa-se que os riscos relativos espaciais brutos (Figuras 2A, 2C e 2E) possuem valores muito próximos dos ajustados (Figuras 2B, 2D e 2F), mudando apenas as áreas de significância.

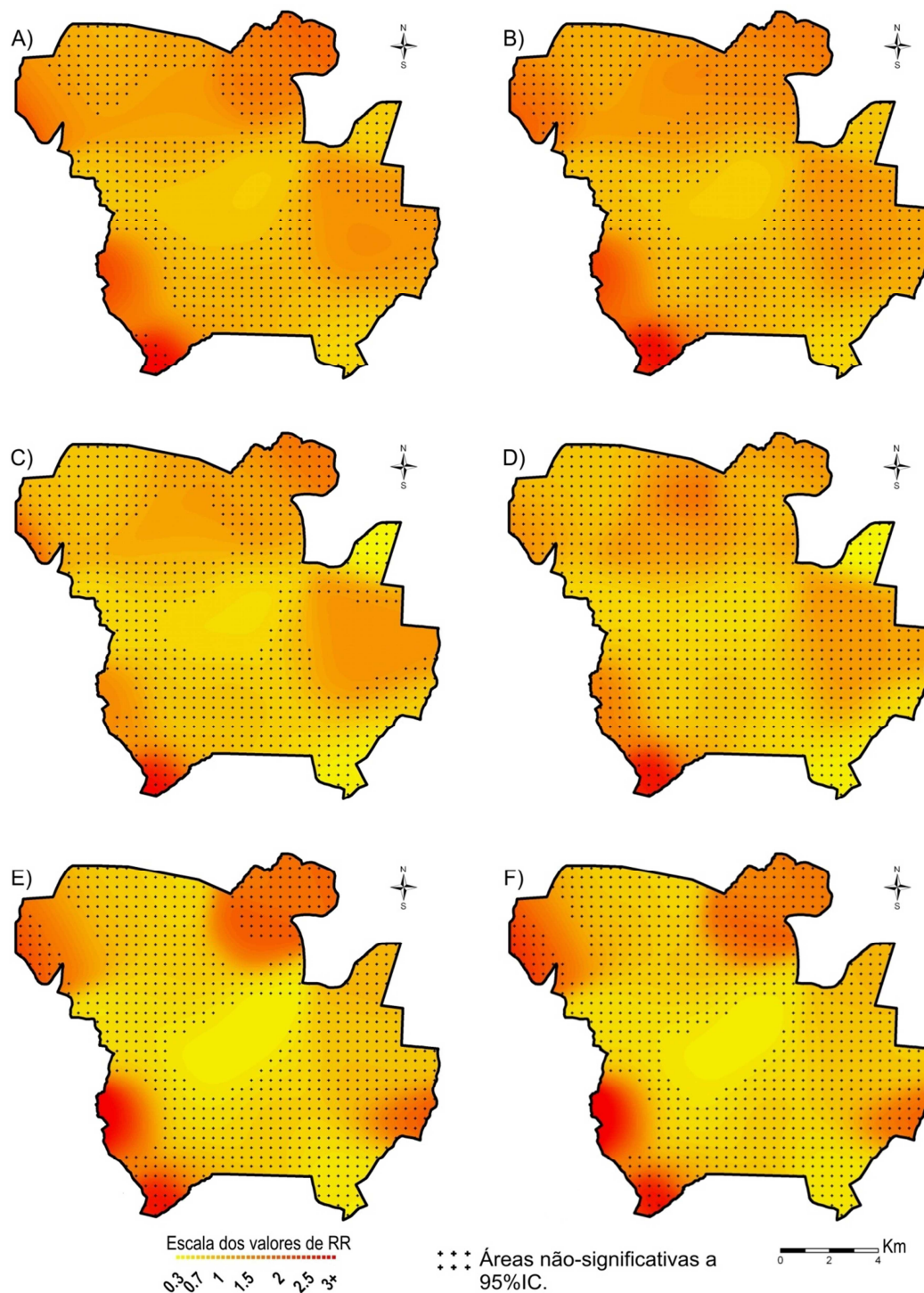
O centro da cidade de Piracicaba destaca-se como uma área de associada à maior proteção nas três análises. As regiões norte, leste e sudoeste estão associadas à maior risco para acidentes entre trabalhadores precarizados, variando a intensidade, a abrangência geográfica e a significância estatística

conforme se compara todos os casos juntos, somente casos leves ou somente casos graves contra os controles.

O ajuste binomial indicou, no modelo bruto, três grandes áreas de risco ao norte ($RR_{m\acute{a}x}=1,8$), leste ($RR_{m\acute{a}x}=1,2$) e sudoeste ($RR_{m\acute{a}x}=2,4$) da região de estudo (Figura 2A). Com a análise multinomial verifica-se que a região ao norte se divide, estando a região noroeste associada à maior risco para casos graves (Figura 2C; $RC_{m\acute{a}x}=1,8$) e a região centro-norte, um distrito industrial chamado UNINORTE, aparece associada à maior risco para casos leves (Figura 2E; $RC_{m\acute{a}x}=1,2$). A área associada à maior risco a leste, também um distrito industrial (UNILESTE), aparece associado à maior risco apenas para casos leves (Figura 2E; $RC_{m\acute{a}x}=1,3$). A região a sudoeste aparece na análise de ambos logitos (leves e graves), porém numa abrangência diferente da análise binomial (Figuras 2C e 2E; $RC_{m\acute{a}x}=3$ para casos graves e $RC_{m\acute{a}x}=1,4$ para casos leves).

Com a inclusão das covariáveis não-espaciais, o risco espacial para casos leves não se mostra mais significativo (Figura 2D), bem como a pequena área a nordeste que antes apresentava risco para casos graves (Figura 2F). Na análise binomial, mostram-se significativos apenas os riscos no centro-oeste e a sudoeste (Figura 2B).

Figura 2: Distribuição espacial do risco relativo (RR) de acidentes de trabalho em Piracicaba, 2007. Risco de acidentes de trabalho (modelo binomial) bruto (A) e ajustado (B). Risco de casos leves em relação aos controles (modelo multinomial) bruto (C) e ajustado (D). Risco de casos graves em relação aos controles (modelo multinomial) bruto (E) e ajustado (F).



4.3 Discussão sobre os resultados dos acidentes de trabalho

No estudo dos fatores associados ao trabalhador, como sexo, idade e escolaridade, os resultados obtidos para os trabalhadores precarizados são semelhantes ao que a literatura indica como associados aos trabalhadores em geral. Observou-se maior proteção conforme o aumento da idade^(45, 53, 54) e da escolaridade^(53, 55) e maior risco para os homens do que para as mulheres^(45, 52, 53, 54).

Ao se verificar possíveis interações, a idade indicou maior proteção para quem possui carteira de trabalho assinada (RC=0,96) do que para quem não a possui (RC=0,98), e ser do sexo masculino corresponde a risco significativo apenas para quem não trabalha na rua (RC=1,96). Na análise multinomial, a idade mostrou-se mais protetora para as mulheres (RC=0,96) do que para os homens (RC=0,97) no estudo dos casos leves e, nos casos graves, foi significativa apenas para as mulheres (RC=0,95). Quanto ao sexo, ser homem indica risco apenas para casos leves entre trabalhadores que não possuem carteira de trabalho assinada (RC=3,2).

Quanto aos fatores relacionados à ocupação, como se o trabalho é na rua, se é doméstico e/ou se é terceirizado, verificou-se um maior risco para trabalhadores domésticos, já indicado na literatura como risco para trabalhadores em geral⁽⁵³⁾ ou apenas em estudo sobre trabalhadoras⁽⁵⁶⁾, mas nesse caso indicam risco significativo apenas para casos graves (RC=2,15). Trabalhar na rua indicou risco de caso leve, para mulheres (RC=1,67).

O tipo de contrato estabelecido na ocupação, já observado na literatura de trabalhadores em geral indicando maior risco para assalariados registrados e funcionários públicos em relação a não-registrados⁽⁵³⁾ e maior incidência de acidentes entre trabalhadores homens com carteira assinada⁽⁵⁵⁾, aparece neste estudo como fator de risco para trabalhadores precarizados que possuem carteira de trabalho assinada, sendo este risco maior entre as mulheres (RC=2,9 para homens e RC=6,3 para mulheres).

A região sudoeste se refere à periferia da cidade de Piracicaba e se caracteriza por uma urbanização descontrolada e ocupada por população carente, talvez por isso esteja associada à maior risco nas três abordagens do estudo.

Os casos leves revelam uma associação com áreas industriais, mas que nesse estudo perdeu a significância estatística após a inclusão de outras covariáveis não-espaciais mostrando que esta associação é explicada pelas outras covariáveis de estudo. Com esse trabalho não se sabe as razões que fizeram das áreas industriais da cidade locais de risco para trabalhadores precarizados, mas sugere-se que estas áreas sejam mais investigadas em trabalhos futuros.

5. Estudo transversal com resposta multinomial

Em estudos transversais faz-se uma amostra aleatória de toda a população de interesse num determinado local e período de tempo, para depois se observar a prevalência de um determinado evento⁽¹⁾. Neste caso, pode-se fazer uso dos modelos ordinais quando se intenta categorizar a resposta de interesse em mais de dois níveis, segundo uma sequência lógica ordinal.

Para ilustrar a aplicação do uso de modelos ordinais em um estudo transversal, utilizaram-se os dados a seguir.

5.1 Criadouros de mosquito no Distrito Sul de Campinas, SP.

As informações sobre criadouros de mosquito utilizadas são provenientes de um estudo caso-controle sobre dengue no Distrito Sul de Campinas⁽⁵⁷⁾. Com o intuito de verificar a prevalência de criadouros de mosquitos na população, consideramos apenas os controles do estudo citado.

Estes controles foram coletados aleatoriamente de um registro universal de domicílios desse local. Através de entrevistas e inspeções peridomiciliares, 727 domicílios foram estudados dentre toda a população (74228 domicílios).

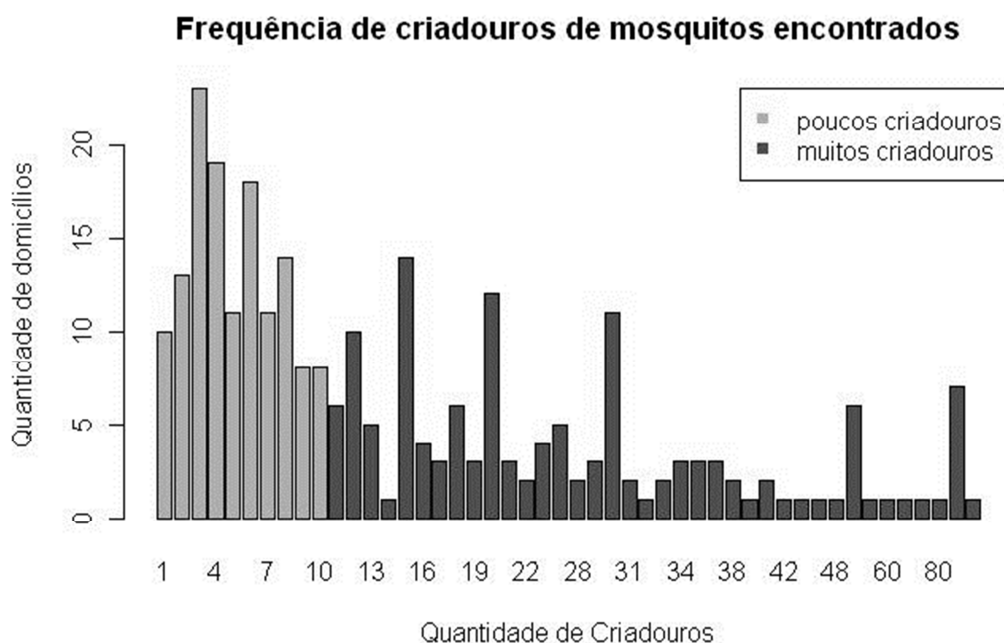
A análise destes dados tem por objetivo determinar a prevalência de criadouros em potencial para o mosquito *Aedes aegypti* na região de estudo, visto que a existência destes criadouros garante ao mosquito a possibilidade de proliferar, propagando assim a dengue^(58, 59). Considerou-se como variável resposta o número de criadouros encontrados no domicílio, o qual foi categorizado em: muitos criadouros ($n_1=136$), poucos criadouros ($n_2=135$) e nenhum criadouro de mosquito ($n_3=456$). O modelo binomial provê medidas que comparam as categorias *encontrou-se* ou *não se encontrou* criadouros no domicílio ou em seu entorno.

A presença de criadouros no domicílio foi categorizada em muitos ou poucos criadouros com o objetivo de destacar áreas que carecem de ações mais urgentes e também direcionar estas ações a públicos distintos: domicílios com

muitos criadouros de mosquitos são, na maioria das vezes, de indivíduos que armazenam material reciclável em suas casas. Os domicílios com poucos criadouros caracterizam-se por moradores com pouca informação sobre a dengue e as formas de prevenção desta.

A categoria *muitos criadouros* indica que foram encontrados mais do que dez criadouros no domicílio ou em seu entorno. A classe *poucos criadouros* indica que foram achados de um a dez possíveis criadouros. Esta classificação foi feita com base na mediana dos números de criadouros do estudo e a Figura 3 traz a distribuição de frequência desta informação.

Figura 3: Distribuição do número de possíveis criadouros encontrados no Distrito Sul de Campinas, 2007.



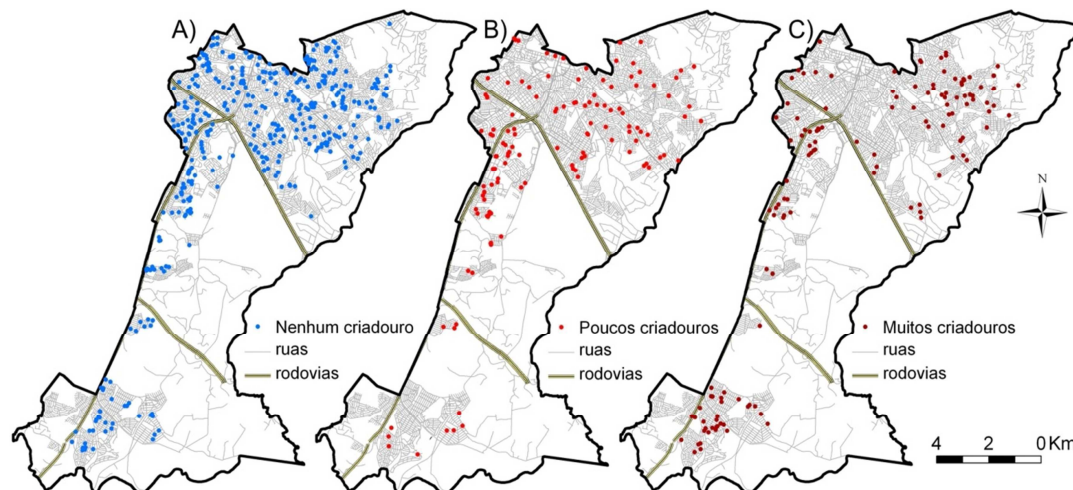
A análise espacial foi feita com base nas coordenadas geográficas dos domicílios. As variáveis não-espaciais analisadas de forma paramétrica neste estudo estão descritas na Tabela 4.

Para a análise destes dados utilizou-se os modelos multinomiais politômico, cumulativo, de razões sequenciais e de categorias adjacentes (seção 3.1.3), além do binomial utilizado para comparação dos resultados.

Tabela 4: Médias e desvios padrões da escolaridade do chefe de domicílio e porcentagens de indivíduos em cada classe por categorias de resposta. Criadouros de mosquitos no Distrito Sul de Campinas, 2007.

| Variáveis | | Mais de dez criadouros (n=136) | De um a dez criadouros (n=135) | Nenhum criadouro (n=456) | Total (N=727) |
|---|------------|-----------------------------------|-----------------------------------|-----------------------------|------------------|
| Escolaridade do chefe do domicílio (anos) | | 5,61 ± 3,86 | 7,62 ± 4,32 | 8,74 ± 4,54 | 7,95 ± 4,54 |
| Saneamento Básico | Rede | 59,56% | 81,48% | 88,60 % | 81,84% |
| | Fossa | 30,15% | 9,63% | 9,43% | 13,34% |
| | Céu Aberto | 10,29% | 8,89% | 1,97% | 4,81% |

Figura 4: Distribuição espacial dos domicílios estudados no Distrito Sul de Campinas, 2007. A) Pontos onde não se encontrou nenhum criadouro. B) Pontos em que se encontrou de um a dez criadouros. C) Pontos onde se encontrou mais de dez criadouros.



A análise multinomial ordinal nos dá cinco diferentes tipos de medidas de risco provenientes de diversas possíveis estruturas de comparações dos logitos. As comparações são as seguintes:

- (1) Muitos criadouros versus pouco ou nenhum criadouro;

- (2) Encontrou-se algum criadouro versus nenhum criadouro;
- (3) Poucos criadouros versus nenhum criadouro;
- (4) Muitos criadouros versus poucos criadouros; e
- (5) Muitos criadouros versus nenhum criadouro.

A primeira comparação é obtida através da estimativa dos primeiros logitos dos modelos cumulativo e de razões sequenciais. A segunda, através do segundo logito do modelo cumulativo, é similar à obtida pelo modelo binomial. Três modelos comparam, no segundo logito, de um a dez criadouros com nenhum criadouro: modelo de razões sequenciais, de categorias adjacentes e o politômico. A quarta comparação é proveniente do modelo de categorias adjacentes e a quinta, do modelo politômico (ambas no primeiro logito).

5.2 Resultados dos criadouros de mosquitos

Todas as estimativas obtidas através das análises paramétricas encontram-se descritas na Tabela 5.

O estudo binomial mostra que a proteção aumenta conforme o aumento da escolaridade do chefe do domicílio (RC=0,93 a cada ano). Também mostra risco aumentado para domicílios com fossa (RC=2,8) e esgoto a céu aberto (RC=5,59) quando comparados com a rede de esgotos.

As análises ordinais apontam para uma maior proteção associada ao aumento da escolaridade do chefe do domicílio e maior risco associado a domicílios sem rede de saneamento básico. O risco de um domicílio que tem fossa, em relação ao que tem rede de esgoto, não foi significativo apenas quando se compara a presença de poucos criadouros contra nenhum criadouro.

As análises dos resíduos dos ajustes GAM multinomiais encontram-se no Apêndice 10.4 e os mapas das análises espaciais estão na seção a seguir.

| | | Escolaridade do Chefe do Domicílio (anos) | | | Saneamento Básico | | | | | | | |
|------------------------|-----------------------|---|--------|--------|-------------------|--------|--------|------------|--------|--------|-------|---------|
| Modelo | Estrutura das RC | RC | LI(RC) | LS(RC) | Fossa | | | Céu Aberto | | | | |
| | | RC | LI(RC) | LS(RC) | RC | LI(RC) | LS(RC) | RC | LI(RC) | LS(RC) | AIC | |
| Modelo Paramétrico | Cumulativo | Muitos vs Poucos ou Nenhum criadouro | 0,87 | 0,82 | 0,91 | 3,88 | 2,41 | 6,23 | 2,84 | 1,36 | 5,92 | 1254,60 |
| | | Muitos ou Poucos vs Nenhum criadouro | 0,92 | 0,88 | 0,95 | 2,10 | 1,35 | 3,29 | 4,55 | 2,07 | 10,03 | |
| | Razões Seqüenciais | Muitos vs Poucos ou Nenhum criadouro | 0,88 | 0,83 | 0,92 | 3,56 | 2,20 | 5,75 | 2,83 | 1,35 | 5,93 | 1259,56 |
| | | Poucos vs Nenhum criadouro | 0,96 | 0,91 | 1,00 | 0,98 | 0,50 | 1,90 | 4,13 | 1,67 | 10,18 | |
| | Categorias Adjacentes | Muitos vs Poucos criadouros | 0,91 | 0,86 | 0,97 | 3,58 | 1,78 | 7,20 | 1,21 | 0,52 | 2,83 | 1259,39 |
| | | Poucos vs Nenhum criadouro | 0,95 | 0,91 | 1,00 | 0,99 | 0,51 | 1,92 | 4,16 | 1,69 | 10,28 | |
| | Politômico | Muitos vs Nenhum criadouro | 0,87 | 0,82 | 0,91 | 3,54 | 2,13 | 5,86 | 5,06 | 2,07 | 12,34 | 1259,39 |
| | | Poucos vs Nenhum criadouro | 0,95 | 0,91 | 1,00 | 0,99 | 0,51 | 1,92 | 4,16 | 1,69 | 10,28 | |
| | Binomial | Muitos ou Poucos vs Nenhum criadouro | 0,92 | 0,88 | 0,95 | 2,13 | 1,36 | 3,33 | 4,55 | 2,06 | 10,03 | 905,43 |
| Modelo Semiparamétrico | Cumulativo | Muitos vs Poucos ou Nenhum criadouro | 0,89 | 0,85 | 0,92 | 4,78 | 3,30 | 6,93 | 3,88 | 2,15 | 7,00 | 1179,69 |
| | | Muitos ou Poucos vs Nenhum criadouro | 0,92 | 0,91 | 0,94 | 2,20 | 1,96 | 2,47 | 2,79 | 2,45 | 3,17 | |
| | Razões Seqüenciais | Muitos vs Poucos ou Nenhum criadouro | 0,89 | 0,86 | 0,92 | 7,43 | 5,58 | 9,90 | 10,88 | 7,87 | 15,05 | 1185,56 |
| | | Poucos vs Nenhum criadouro | 0,95 | 0,92 | 0,99 | 1,51 | 0,92 | 2,46 | 5,71 | 3,68 | 8,85 | |
| | Categorias Adjacentes | Muitos vs Poucos criadouros | 0,92 | 0,88 | 0,95 | 6,23 | 4,61 | 8,43 | 5,02 | 3,64 | 6,94 | 1195,22 |
| | | Poucos vs Nenhum criadouro | 0,95 | 0,92 | 0,98 | 1,52 | 0,93 | 2,50 | 6,17 | 1,97 | 9,57 | |
| | Politômico | Muitos vs Nenhum criadouro | 0,88 | 0,85 | 0,91 | 8,21 | 6,17 | 10,92 | 25,15 | 18,36 | 34,45 | 1183,96 |
| | | Poucos vs Nenhum criadouro | 0,95 | 0,92 | 0,99 | 1,49 | 0,92 | 2,44 | 5,81 | 3,75 | 8,99 | |
| | Binomial | Muitos ou Poucos vs Nenhum criadouro | 0,93 | 0,91 | 0,95 | 2,80 | 2,27 | 3,45 | 5,59 | 4,49 | 6,95 | 883,3 |

AIC: Critério de Informação de Akaike; vs: versus.

RC: Razão de Chances; LI: 95% Limite Inferior; LS: 95% Limite Superior.

Tabela 5: Razões de chances obtidas através do ajuste dos modelos linear generalizado (paramétrico) e aditivo generalizado (semiparamétrico) para os dados de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007 (N=727).

Análises espaciais

A Figura 5 mostra as análises espaciais obtidas através de modelos semiparamétricos para os ajustes multinomiais e para o binomial.

Ao se observar os riscos espaciais relacionados a muitos criadouros nota-se duas áreas de maior risco significativo, uma a nordeste (NE) e outra a leste (L) e uma área de proteção no centro da região norte (N). Estas áreas podem ser notadas ao se estudar o risco de muitos criadouros versus pouco ou nenhum criadouro que no modelo cumulativo chega a $RR_{\text{máx}}(\text{NE})=1,2$, $RR_{\text{máx}}(\text{L})=2,8$ e $RR_{\text{máx}}(\text{N})=0,6$ (Figura 5A) e no modelo de razões sequenciais tem-se os valores $RR_{\text{máx}}(\text{NE})=4$, $RR_{\text{máx}}(\text{L})=1,6$ e o $RR_{\text{máx}}(\text{N})=0,2$ (Figura 5C). Pode-se ainda estudar o risco de encontrar domicílios com muitos criadouros comparado ao risco de se encontrar poucos criadouros o qual é ajustado pelo modelo de categorias adjacentes e pode ser visto na Figura 5E em que observa-se os valores $RR_{\text{máx}}(\text{NE})=4$, $RR_{\text{máx}}(\text{L})=2,2$ e $RR_{\text{máx}}(\text{N})=0,2$ ou pode-se comparar este risco ao de não encontrar nenhum criadouro, como se vê na Figura 5G do modelo politômico onde observa-se que $RR_{\text{máx}}(\text{NE})=6$, $RR_{\text{máx}}(\text{L})=1,2$ e $RR_{\text{máx}}(\text{N})=0,2$.

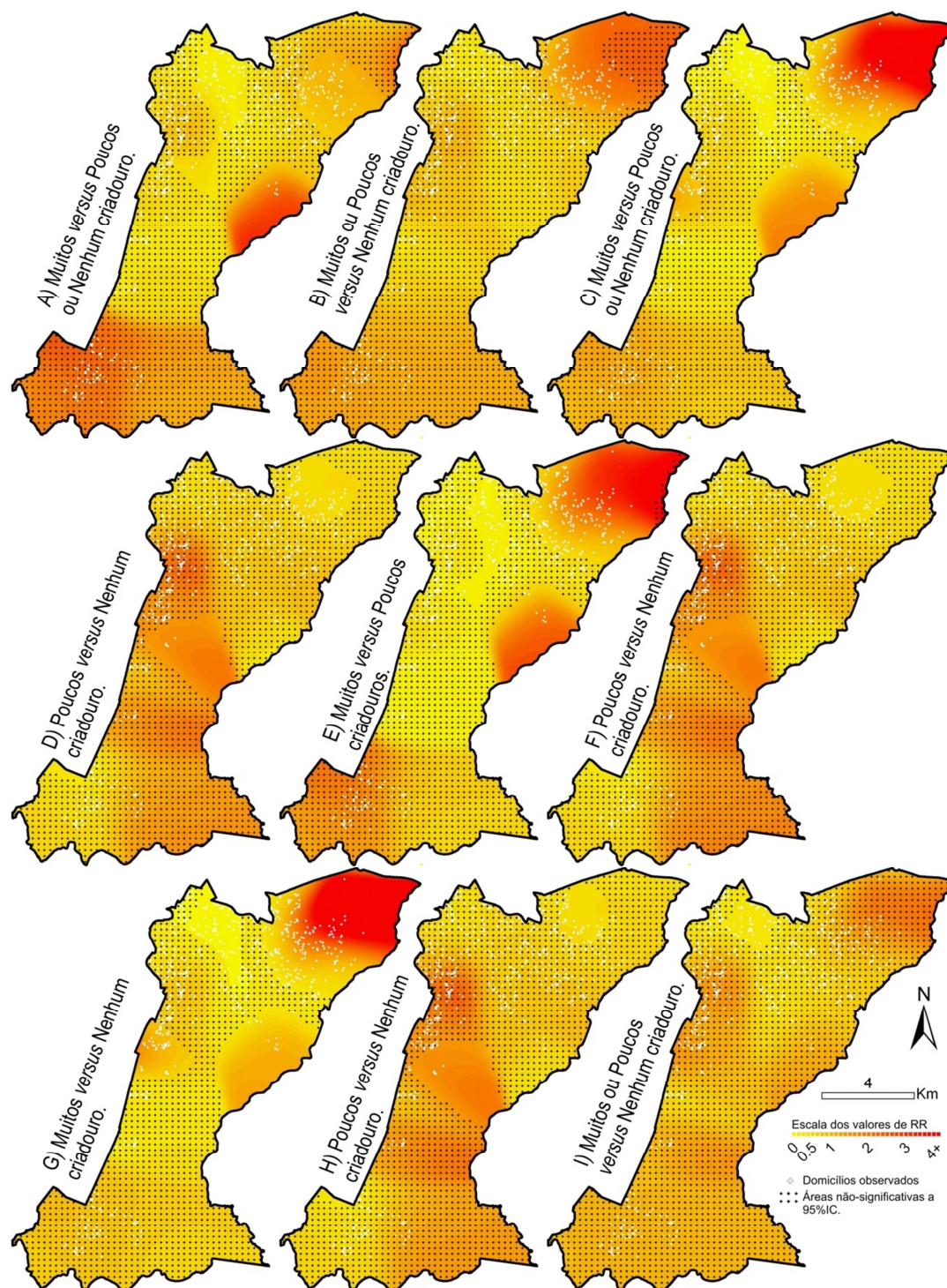
A Figura 5B provê uma análise espacial similar à da análise binomial, pois neste logito do modelo cumulativo tem-se o risco de muitos ou poucos criadouros versus nenhum criadouro, e vê-se uma área de risco significativo ao norte ($RR_{\text{máx}}=2,3$).

Para o estudo do risco de se encontrar poucos criadouros em relação a nenhum criadouro, três modelos provêm esta medida, que são os modelos de razões sequenciais (Figura 5D), de categorias adjacentes (Figura 5F) e o politômico (Figura 5H). Nestas análises observa-se uma área associada à maior proteção ao norte ($RR_{\text{máx}}=0,6$) e outra, associada à maior risco, no centro da região de estudo ($RR_{\text{máx}}=2$).

A Figura 5G ainda mostra uma área em que as estimativas foram significativas e mostraram uma associação de risco no centro-oeste da região de estudo ($RR_{\text{máx}}=1,4$), a qual não havia aparecido nas demais análises.

A análise binomial destacou apenas uma pequena área associada à maior proteção no centro da região norte do distrito estudado (Figura 5I, $RC_{\text{máx}}=0,3$).

Figura 5: Distribuição espacial do risco relativo (RR) ajustado de criadouros de mosquitos em potencial no Distrito Sul de Campinas, 2007. Modelos multinomiais cumulativo (A e B), de razões sequenciais (C e D), de categorias adjacentes (E e F), politômico (G e H) e modelo binomial (I).



5.3 Discussão sobre os resultados dos criadouros de mosquitos

Em estudos transversais em que a variável resposta de interesse segue uma distribuição de probabilidade multinomial e que as categorias desta resposta seguem uma ordem lógica é preciso escolher qual dos modelos ordinais ajustará melhor os dados. Nesse estudo utilizaram-se quatro modelos logísticos - o politômico, o de categorias adjacentes, o de razões sequenciais e o cumulativo -, entretanto isso só foi realizado para exemplificar as variadas estruturas de logitos e comparar os resultados encontrados com cada uma delas.

Para que um pesquisador escolha qual destes modelos lhe servirá melhor, é preciso observar se seus dados permitem que se acumulem categorias ou não durante o ajuste.

Se o objetivo do estudo é observar o risco associado a categorias específicas desta resposta, então o ideal é não acumula-las e, neste caso, sugere-se dois modelos: o politômico e o de categorias adjacentes. Destes dois, o politômico é a escolha ideal quando se deseja fixar uma das categorias como referência para todas as demais comparações para, por exemplo, se ter resultados comparáveis com o que já existe na literatura; caso contrário, o de categorias adjacentes é mais indicado porque compara entre si situações mais similares.

Se, por outro lado, não se tem objeções quanto ao acúmulo de categorias, o modelo logístico cumulativo é a melhor escolha porque, ao utilizar todos os dados no ajuste de cada logito, provê estimativas mais consistentes, razões de chances com intervalos de confiança menos amplos e análises espaciais com menores efeitos de borda.

O modelo de razões de chance é uma opção que visa equilibrar a escolha entre um modelo consistente, porque também acumula algumas categorias, e a provisão de riscos específicos para uma categoria da variável resposta, se assimilando ao modelo de categorias adjacentes.

No estudo dos criadouros de mosquitos observou-se que o modelo cumulativo foi o que proveu o menor valor do Critério de Informação de Akaike (AIC) dentre os modelos multinomiais, indicando que foi o mais bem ajustado. Ao se inserir o ajuste do risco espacial nos modelos de razões sequenciais, de

categorias adjacentes e o politômico, estes mostraram certa instabilidade nas estimativas das razões de chances. Observa-se tal instabilidade nos intervalos de confiança associados à variável saneamento básico, pois ficarem muito amplos (o risco para domicílios com esgoto a céu aberto variou de 7,87 a 15,05 no modelo de razões sequenciais; de 1,97 a 9,57 no de categorias adjacentes; e de 18,36 a 34,45 no politômico).

Entretanto, se o objetivo desta análise fosse encontrar riscos específicos de cada categoria de resposta sem acumulá-las, o modelo que melhor ajustou esses dados foi o de categorias adjacentes. Apesar de indicar o maior valor de AIC, este modelo proveu as estimativas de razões de chance mais consistentes e as análises espaciais de risco foram semelhantes às dos modelos de razões sequenciais e politômico.

Pode-se notar que, mesmo com várias opções de modelos, quando se confrontam os logitos que a mesma estrutura de comparação, os valores estimados são bem próximos.

Por não se encontrar outros estudos específicos sobre possíveis criadouros de mosquitos, não se pode comparar os resultados encontrados aqui com a literatura, mas pode-se notar que, na análise paramétrica, todas as comparações dos modelos multinomiais seguiram a mesma direção que os resultados encontrados na análise binomial, porém com diferentes magnitudes.

Os modelos indicam proteção associada a aumento da escolaridade do chefe domiciliar e risco associado a não possuir rede de saneamento básico. Entretanto, a proteção para a escolaridade é maior nas comparações com domicílios em que se encontraram muitos possíveis criadouros e o risco associado às casas com fossa em relação às com rede de saneamento básico, quando comparados muitos contra poucos criadouros não foi significativo.

Apenas 35 domicílios do estudo possuíam esgoto a céu aberto⁽⁵⁷⁾, algo que resulta em estimativas com grande variabilidade (intervalos de confiança amplos), principalmente nas análises multinomiais, em que há também a estratificação da variável resposta.

Nas análises espaciais as áreas com riscos significativos aparecem apenas nas análises multinomiais, as quais apontam regiões específicas de ação: ao centro e a noroeste do Distrito Sul de Campinas.

6. Discussão Geral

Este trabalho apresentou uma série de maneiras de se estimar a distribuição espacial do RR, ajustada a outras covariáveis do estudo, quando a resposta é multinomial, seja nominal ou ordinal. A escolha por uma das análises propostas é discutida na literatura^(32, 33) e a seção 5.3 a complementa para o caso específico dos estudos epidemiológicos. Os pesquisadores que desejarem fazer uso destas ferramentas ainda precisam buscar selecionar a alternativa que melhor corresponde a seus pressupostos e objetivos iniciais.

O fato de neste trabalho terem sido desenvolvidas todas as funções necessárias para a análise espacial de dados epidemiológicos com resposta multinomial em um *software*-livre permite que outros epidemiologistas apliquem este método a seus estudos, quando apropriado.

Por exemplo, Shimakura et al.⁽¹¹⁾ analisam a distribuição espacial da mortalidade infantil em Porto Alegre através de mapas específicos para mortalidade infantil neonatal e pós-neonatal e comparando-as à população sob risco no período em que ocorreram os óbitos. Esse é um caso em que se poderia ter utilizado o ajuste de um GAM multinomial politômico, ao invés de se ajustarem dois GAM binomiais, como os pesquisadores fizeram.

O estudo de coorte de Ballester et al.⁽⁴⁾, poderia subclassificar a poluição do ar em diferentes níveis de exposição e utilizar um GAM ordinal em sua análise espacial, tornando as informações que obteve mais específicas.

Os estudos sobre a distribuição espacial da dengue conduzidos por Braga et al.⁽⁶⁾ e Honório et al.⁽⁹⁾ poderiam ter subclassificado a dengue através de seus diferentes sorotipos e Carvalho et al.⁽⁷⁾, que estudou a distribuição espacial de traumas dentais, poderia especificá-los subdividindo a resposta por tipo de trauma dental e utilizar um GAM politômico para a análise espacial.

Quanto às discussões específicas sobre os resultados das análises de dados de estudo caso-controle e de estudo transversal, estas se encontram nas seções 4.3 e 5.3, respectivamente.

Destaca-se a importância da precaução no uso de processos de pós-estratificação, que ocasionalmente resultam em algumas categorias com um

número de indivíduos não representativo ou mesmo em amplos intervalos de confiança, seja para a análise paramétrica ou para a espacial. Vê-se como exemplo as análises paramétricas do esgoto a céu aberto, no estudo de criadouros de mosquitos (seção 5.3), e no estudo da interação entre possuir carteira de trabalho e ter direitos trabalhistas, no estudo dos acidentes de trabalho (seção 4.3).

Sugere-se, em estudos onde se tem por objetivo considerar um desfecho com mais de duas categorias, incluírem no planejamento destes o tipo de análise que se deseja fazer. Isso facilitaria na escolha de um desenho de estudo adequado ao tipo de análise (ordinal ou com um grupo de referência) e se evitaria os problemas de pós-estratificação da variável resposta, pois pode haver um planejamento amostral que garanta um número mínimo de indivíduos em cada categoria de resposta.

Em relação às análises dos dados reais de acidentes de trabalho e de criadouros de mosquitos, espera-se contribuir para futuras ações de intervenção na área da saúde. Em razão dos mapas apresentarem áreas específicas de risco para cada tipo de desfecho, as ações podem ser orientadas de forma a estabelecer prioridades, tendo-se a opção de escolher áreas em que o risco da doença foi mais alto ou áreas de risco para casos mais graves. Pode-se também agir de acordo com a realidade de cada área, como no caso dos trabalhadores, a ação em distritos industriais pode ser diferente da ação que se planeja para uma área comercial da cidade.

7. Conclusão

Os mapas de risco analisados nos Capítulos 4 e 5 ilustram a riqueza das análises espaciais multinomiais no entendimento do fenômeno estudado. Pois, tanto no estudo dos dados de acidentes de trabalho, quanto no estudo de possíveis criadouros de mosquito, os resultados obtidos com as análises espaciais multinomiais trouxeram informações diferentes das obtidas somente pela análise multinomial paramétrica ou pela análise espacial binomial, considerados métodos usuais.

A análise multinomial, ainda que se observarmos somente os modelos paramétricos, permite formular hipóteses mais precisas do que apenas com as análises binomiais. Nos estudos dos acidentes de trabalho, no modelo multinomial aparecem novas relações que no binomial não haviam surgido, como o risco para quem não trabalha em ambiente doméstico e algumas novas interações. Para os criadouros de mosquitos, alterou-se a grandeza dos valores de risco e proteção encontrados com o modelo binomial.

Quanto à distribuição do risco no espaço, no estudo dos acidentes de trabalho, a análise espacial multinomial especificou a que tipo de caso as áreas de risco, primeiramente encontradas no modelo binomial, estavam associadas. No estudo dos criadouros, tal análise destacou áreas que no estudo dicotômico não haviam aparecido, como as áreas de risco ao centro e a noroeste da região de estudo.

Esse trabalho cobre todos os objetivos iniciais e apresenta de forma completa e minuciosa o método desenvolvido e a forma de aplicá-lo e interpretar seus resultados, bem como discute a adequação de cada um a alguns dos estudos epidemiológicos.

8. Referências Bibliográficas

1. Rouquayrol MZ, Almeida Filho N. Introdução a epidemiologia moderna, 6a edição. Rio de Janeiro: MEDSI, 2003.
2. Medronho RA. Epidemiologia. Rio de Janeiro: Editora Atheneu, 2002.
3. Hastie TJ, Tibshirani RJ. Generalized additive models. London: Chapman and Hall., 1990.
4. Ballester F, Estarlich M, Iñiguez C, Llop S, Ramón R, Esplugues A, Lacasaña M, Rebagliato M. Air pollution exposure during pregnancy and reduced birth size: a prospective birth cohort study in Valencia, Spain. *Environ Health*. 2010 ;9:6.
5. Bayentin L, El Adlouni S, Ouarda TB, Gosselin P, Doyon B, Chebana F. Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. *Int J Health Geogr*. 2010 ;9:5.
6. Braga C, Luna CF, Martelli CM, de Souza WV, Cordeiro MT, Alexander N, de Albuquerque Mde F, Júnior JC, Marques ET. Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. *Acta Trop*. 2010 ;113(3):234-40.
7. Carvalho ML, Moysés SJ, Bueno RE, Shimakura S, Moysés ST. A geographical population analysis of dental trauma in school-children aged 12 and 15 in the city of Curitiba-Brazil. *BMC Health Serv Res*. 2010 ;10:203.
8. French JL, Wand MP. Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics* 2004, 5(2):177-191.
9. Honório NA, Nogueira RM, Codeço CT, Carvalho MS, Cruz OG, Magalhães Mde A, de Araújo JM, de Araújo ES, Gomes MQ, Pinheiro LS, da Silva Pinel C, Lourenço-de-Oliveira R. Spatial evaluation and modeling of Dengue seroprevalence and vector density in Rio de Janeiro, Brazil. *PLoS Negl Trop Dis*. 2009 ;3(11):e545.
10. Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A. Cluster detection methods applied to the Upper Cape Cod cancer data. *Environ. Health*. 2005; 4:19.
11. Shimakura SE, Carvalho MS, Aerts DRGC, Flores R. Distribuição espacial do

- risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. *Cad. Saúde Pública* 2001; 17(5):1251-1261.
12. Stephan C, Henn CA, Donalisio MR. Geographic expression of AIDS epidemic in Campinas, Southeastern Brazil, between 1980 and 2005. *Rev Saúde Pública*. 2010 ;44(5):812-9.
 13. Vieira V, Webster T, Aschengrau A, Ozonoff D. A method for spatial analysis of risk in a population-based case-control study. *Intern. J. of Hygiene and Environmental Health*. 2002;205(1-2):115-120.
 14. Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: An application of generalized additive models to case-control data. *Environmental Health: A Global Access Science Source* 2005; 4:11.
 15. Vieira VM, Hart JE, Webster TF, Weinberg J, Puett R, Laden F, Costenbader KH, Karlson EW. Association between residences in U.S. northern latitudes and rheumatoid arthritis: A spatial analysis of the Nurses' Health Study. *Environ Health Perspect*. 2010 ;118(7):957-61.
 16. Webster T, Vieira V, Weinberg J, Aschengrau A. Method for mapping population-based case-control studies: an application using generalized additive models. *International Journal of Health Geographics* 2006, 5:26.
 17. Young RL, Weinberg J, Vieira V, Ozonoff A, Webster TF. A power comparison of generalized additive models and the spatial scan statistic in a case-control setting. *Int J Health Geogr*. 2010 ;9:37.
 18. Zangirolani LT, Cordeiro R, de Medeiros MA, Stephan C. Spatial distribution of risks for work-related injuries in a city of Southeastern Brazil. *Rev. Saúde Pública*. 2008; 42(2): 287-93.
 19. Abreu MNS, Siqueira AL, Cardoso CS, Caiaffa WT. Ordinal logistic regression models: application in quality of life studies. *Cadernos de Saúde Pública* 2008;24(Sup 4):S581-S591.
 20. Campbell PT, Sloan M, Kreiger N. Utility of proxy versus index respondent information in a population-based case-control study of rapidly fatal cancers. *Annals of Epidemiology*. 2007;17(4):253-7.
 21. Carrière I, Bouyer J. Random-effect models for ordinal responses: application to self-reported disability among older persons. *Revue d'Épidémiologie et de*

Santé Publique, 2006; 54(1):61-72.

22. Chiu BC, Dave BJ, Blair A, Gapstur SM, Chmiel JS, Fought AJ, Zahm SH, Weisenburger DD. Cigarette smoking, familial hematopoietic cancer, hair dye use, and risk of t(14;18)-defined subtypes of non-Hodgkin's lymphoma. *American Journal of Epidemiology*. 2007;165(6):652-9.
23. Ferreira V, Brito C, Portela M, Escosteguy C, Lima S. DOTS in primary care units in the city of Rio de Janeiro, Southeastern Brazil. *Revista de Saúde Pública*, 2011;45(1):40-8.
24. Haelterman E, Marcoux S, Croteau A, Dramaix M. Population-based study on occupational risk factors for preeclampsia and gestational hypertension. *Scand J Work Environ Health*. 2007;33(4):304-17.
25. Kazembe LN, Namangale JJ. A Bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. *European Journal of Epidemiology*, 2007 ;22(8):545-56.
26. Kim H, Friedlander Y, Longstreth WT Jr, Edwards KL, Schwartz SM, Siscovick DS. Family history as a risk factor for stroke in young women. *American Journal of Preventive Medicine*. 2004;27(5):391-6.
27. Lopes C, Andreozzi VL, Ramos E, Sá Carvalho M. Modelling over week patterns of alcohol consumption. *Alcohol and Alcoholism*, 2008 ;43(2):215-22.
28. Mathes RW, Malone KE, Daling JR, Porter PL, Li CI. Relationship between histamine2-receptor antagonist medications and risk of invasive breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2008;17(1):67-72.
29. Shrubsole MJ, Wu H, Ness RM, Shyr Y, Smalley WE, Zheng W. Alcohol drinking, cigarette smoking, and risk of colorectal adenomatous and hyperplastic polyps. *American Journal of Epidemiology*. 2008;167(9):1050-8.
30. Tan Q, Christiansen L, Christensen K, Kruse TA, Bathum L. Apolipoprotein E genotype frequency patterns in aged Danes as revealed by logistic regression models. *European Journal of Epidemiology*. 2004;19(7):651-6.
31. Yip PS, Lau EH, Lam KF, Huggins RM. A chain multinomial model for estimating the real-time fatality rate of a disease, with an application to severe acute respiratory syndrome. *American Journal of Epidemiology*, 2005 ;161(7):700-6.
32. Agresti A. *Categorical data analysis*. 2nd ed. New York: John Wiley & Sons;

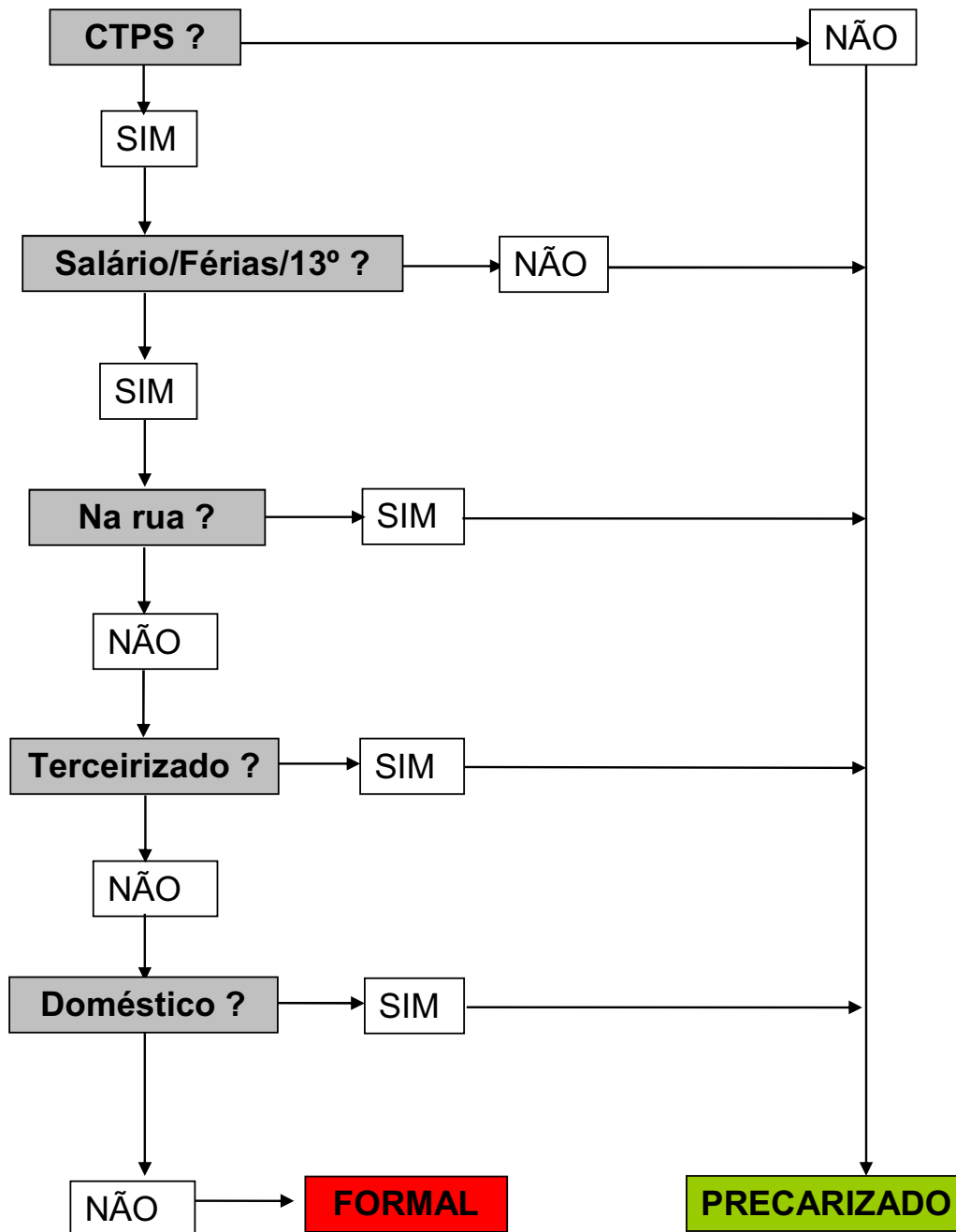
2002.

33. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* 1997;26:1323-1333.
34. Engel J. Polytomous logistic regression. *Statistica Neerlandica* 1988;42(4):233-252.
35. Feinberg B. Analysis of cross-classified data. 2nd ed. Cambridge: Massachusetts Institute of Technology Press; 1980.
36. Mafra ACCN, Nucci LB, Cordeiro R, Stephan C. Estudo caso-controle com resposta multinomial: uma proposta de análise. *Cadernos de Saúde Pública* 2010; 26(3):451-459.
37. Nelder JA, Wedderburn RWM. Generalized Linear Models. *J. R. Stat. Soc. A.* 1972, vol. 135, p. 370-384.
38. Kelsall JE, Diggle PJ. Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics* 1998;47:559-73.
39. Paulino CD, Singer JM. *Análise de Dados Categorizados*. São Paulo: Blucher, 2006.
40. Hosmer Jr, DW, Lemeshow, S. *Applied logistic regression*. 2000. New York, John Wiley & Sons.
41. Wood, SN. *Generalized Additive Models: an introduction with R*. Boca Raton: Chapman & Hall/CRC, 2006.
42. Bithell J. An application of density estimation to geographical epidemiology. *Statistics in Medicine* 1990;9:691-701.
43. Diggle PJ. *Statistical analysis of spatial point patterns*, 2nd ed. London: Arnold, 2003.
44. DeGroot MH. *Probability and statistics*, 2nd ed. Reading, MA: Addison-Wesley Publishing Company, 1986.
45. Bailey TC, Cordeiro R, Lourenço RW. Semiparametric Modeling of the Spatial Distribution of Occupational Accident Risk in the Casual Labor Market, Piracicaba, Southeast Brazil. *Risk Analysis* 2007;27(2):421-431.
46. Kelsall JE, Diggle PJ. Kernel estimation of relative risk. *Statistics in Medicine* 1995; 1(1-2):03-16.
47. The R Project for Statistical Computing [programa de computador]. Disponível

em <http://www.r-project.org/> . Acessado em 22/07/2011.

48. Página do Laboratório de Análise Espacial de Dados Epidemiológicos. Disponível em: <http://www.epigeo.net/> . Acessado em 22/07/2011.
49. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia, PA: Lippincott-Raven, 1998.
50. Cordeiro R. O mito da doença rara. Revista Brasileira de Epidemiologia 2005;8:111-6.
51. Stephan C. Distribuição do risco de acidente do trabalho entre trabalhadores precarizados de Piracicaba. [Tese de Doutorado]. Campinas (SP): Faculdade de Ciências Médicas, Universidade Estadual de Campinas, 2008.
52. Cordeiro R, Vilela RAG, Medeiros MAT, Gonçalves CGO, Bragantini CA, Varolla AJ et al. O sistema de vigilância de acidentes do trabalho de Piracicaba. Cadernos de Saúde Pública 2005; 21:1574-83.
53. Barata RCB, Ribeiro MCSA, Moraes JC. Acidentes de trabalho referidos por trabalhadores moradores em área urbana no interior do estado de São Paulo em 1994. Informe Epidemiológico do SUS 2000; 9:199-210.
54. Cordeiro R, Prestes SCC, Clemente APG, Diniz CS, Sakate MS, Donalisio MR. Incidência de acidentes do trabalho não-fatais em localidade do Sudeste do Brasil. Cadernos de Saúde Pública 2006; 22(2): 387-393.
55. Santana VS, Maia AP, Carvalho C, Luz G. Acidentes de trabalho não fatais: diferenças de gênero e tipo de contrato de trabalho. Cadernos de Saúde Pública 2003; 19:481-93.
56. Santana VS, Amorim AM, Oliveira R, Xavier S, Iriart J, Belitardo L. Emprego em serviços domésticos e acidentes de trabalho não fatais. Revista de Saúde Pública 2003; 37:65-74.
57. Andrade VR. Distribuição espacial do risco de dengue em região do município de Campinas [Tese de Doutorado]. Campinas (SP): Faculdade de Ciências Médicas, Universidade Estadual de Campinas, 2009.
58. Pontes, RJ, Freeman, J, Oliveira-Lima, JW, Hodgson, JC, Spielman, A. Vector densities that potentiate dengue outbreaks in a brazilian city. American Journal of Tropical Medicine and Hygiene. 2000; 62: 378–383.
59. Teixeira, MG, Costa, MCN, Barreto, F, Barreto, ML. Dengue: twenty-five years since reemergence in brazil. Cadernos de Saúde Pública. 2009; 25: S7–S18.

9. Anexo: roteiro de identificação de trabalho precarizado



Extraído de: Stephan C. Distribuição do risco de acidente do trabalho entre trabalhadores precarizados de Piracicaba. [Tese de Doutorado]. Campinas (SP): Faculdade de Ciências Médicas, Universidade Estadual de Campinas, 2008.

10. Apêndices

10.1 Obtenção dos valores de p , z e w para o ajuste dos Modelos Aditivos Generalizados com resposta multinomial

Foi necessário obter, para cada um dos modelos multinomiais propostos, medidas definidas em Hastie e Tibshirani⁽³⁾ (página 138) e tidas como indispensáveis para este tipo de ajuste dos modelos aditivos generalizados. Tais medidas estão apresentadas no Quadro 1, no item 4.3 do Capítulo 4 desta tese, mas foram obtidas como se segue abaixo.

Parte-se da definição da variável dependente ajustada (z) e dos pesos (w) definidos como $z_i = \eta_i^0 + (y_i - \mu_i^0) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_0$ e $w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)_0^2 V_i^0$, respectivamente, sendo V a variância de Y em μ e η , o preditor linear de cada modelo.

Aqui será mostrada também a forma de obtenção das probabilidades à partir dos logitos estimados em cada modelo.

Para se ter acesso à álgebra necessária à obtenção de cada uma destas medidas, optou-se por demonstrá-las como se segue:

Modelo Polîtômico (ou polytomous logistic model):

Neste caso, considere o preditor linear:

$$\eta_j = \ln \left(\frac{p_j}{p_r} \right), \quad j = 1, 2, \dots, r-1$$

onde p_j é a probabilidade de um determinado evento ter resposta j , dadas as covariáveis incluídas no modelo.

Daí, tem-se que a derivada de η_j em relação a p_j é dada por:

$$\frac{\partial \eta_j}{\partial p_j} = \frac{p_r}{p_j} \times \frac{p_r}{p_r^2} = \frac{1}{p_j}, \quad j = 1, 2, \dots, r-1$$

Dessa forma,

$$z_{ij} = \eta_{ij}^0 + (y_{ij} - p_{ij}^0) \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0 = \eta_j + \frac{(y_{ij} - p_j)}{p_j}, \quad j = 1, 2, \dots, r-1$$

$$E w_{ij}^{-1} = \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0^2 V_{ij}^0 = \left(\frac{1}{p_j} \right)^2 p_j (1 - p_j) = \frac{(1 - p_j)}{p_j}, \text{ por tanto,}$$

$$w_j = \frac{1}{w_j^{-1}} = \frac{p_j}{(1 - p_j)}, \quad j = 1, 2, \dots, r-1$$

Para a obtenção dos p_j s parte-se do pressuposto de que $\sum_{j=1}^r p_j = 1$ e, desta igualdade tem-se que $\sum_{m=1}^{r-1} p_m = 1 - p_r$ (i).

$$\text{Como } \exp(\eta_j) = \frac{p_j}{p_r}, \quad j = 1, 2, \dots, r-1, \text{ temos que } \sum_{m=1}^{r-1} \exp(\eta_m) = \sum_{m=1}^{r-1} \frac{p_m}{p_r} = \frac{\sum_{m=1}^{r-1} p_m}{p_r},$$

que substituindo (i) nos dá:

$$\sum_{m=1}^{r-1} \exp(\eta_m) = \frac{1 - p_r}{p_r}.$$

Isolando p_r , temos:

$$p_r = \frac{1}{1 + \sum_{m=1}^{r-1} \exp(\eta_m)}.$$

À partir de p_r , para obter os demais valores de p , sabendo que $\exp(\eta_j) = \frac{p_j}{p_r}$,

$j = 1, 2, \dots, r-1$, basta:

$$p_j = \exp(\eta_j) \times p_r, \quad j = 1, 2, \dots, r-1.$$

Modelo Cumulativo (ou cumulative model):

Para esse modelo temos o preditor linear:

$$\eta_j = \ln \left(\frac{\sum_{i=1}^j p_i}{\sum_{i=j+1}^r p_i} \right), \quad j = 1, 2, \dots, r-1$$

onde p_j é a probabilidade de um determinado evento ter resposta j , dadas as covariáveis incluídas no modelo.

Daí, tem-se que a derivada de η_j em relação a p_j é dada por:

$$\frac{\partial \eta_j}{\partial p_j} = \frac{\sum_{l=j+1}^r p_l}{\sum_{i=1}^j p_i} \times \frac{\sum_{l=j+1}^r p_l}{\left(\sum_{l=j+1}^r p_l \right)^2} = \frac{1}{\sum_{i=1}^j p_i}, \quad j = 1, 2, \dots, r-1$$

Dessa forma,

$$z_{ij} = \eta_{ij}^0 + (y_{ij} - p_{ij}^0) \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0 = \eta_j + \frac{(y_j - p_j)}{\sum_{i=1}^j p_i}, \quad j = 1, 2, \dots, r-1$$

$$w_{ij}^{-1} = \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0^2 V_{ij}^0 = \left(\frac{1}{\sum_{i=1}^j p_i} \right)^2 p_j (1 - p_j) = \frac{p_j (1 - p_j)}{\left(\sum_{i=1}^j p_i \right)^2}$$

E, portanto,

$$w_i = \frac{1}{w_i^{-1}} = \frac{\left(\sum_{i=1}^j p_i \right)^2}{p_j (1 - p_j)}, \quad j = 1, 2, \dots, r-1.$$

Para a obtenção dos p_j s novamente parte-se do pressuposto de que $\sum_{j=1}^r p_j = 1$

, porém, desta igualdade pode-se ter que $\sum_{m=2}^r p_m = 1 - p_1$ (ii).

Como $\eta_1 = \ln \left(\frac{p_1}{\sum_{m=2}^r p_m} \right)$, temos que, $\exp(\eta_1) = \frac{p_1}{\sum_{m=2}^r p_m}$ que substituindo (ii) nos dá:

$\exp(\eta_1) = \frac{p_1}{1 - p_1}$. Isolando-se o p_1 , tem-se que:

$$p_1 = \frac{\exp(\eta_1)}{1 + \exp(\eta_1)}.$$

$$\eta_2 = \ln \left(\frac{p_1 + p_2}{\sum_{m=3}^r p_m} \right) \text{ e sendo } \sum_{m=3}^r p_m = 1 - p_1 - p_2 \quad (\text{iii}), \text{ para obter } p_2, \text{ basta conhecer}$$

p_1 e η_2 , e ao isolar p_2 , tem-se:

$$p_2 = \frac{\exp(\eta_2)}{1 + \exp(\eta_2)} - p_1$$

À partir de p_1 , pode-se obter os demais valores de p de forma sequencial, semelhante ao p_2 :

$$p_j = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} - \sum_{m=1}^{j-1} p_m, \quad j = 2, \dots, r-1.$$

Tendo todas as $r-1$ probabilidades, obtém-se p_r da seguinte maneira:

$$\text{Sendo } \eta_{r-1} = \ln \left(\frac{\sum_{m=1}^{r-1} p_m}{p_r} \right), \text{ temos que } p_r = \frac{\sum_{m=1}^{r-1} p_m}{\exp(\eta_{r-1})}.$$

Modelo de Razões Contínuas (ou continuation-ratio model):

O preditor linear destes modelo é:

$$\eta_j = \ln \left(\frac{p_j}{\sum_{l=j+1}^r p_l} \right), \quad j = 1, 2, \dots, r-1$$

onde p_j é a probabilidade de um determinado evento ter resposta j , dadas as covariáveis incluídas no modelo.

Daí, tem-se que a derivada de η_j em relação a p_j é dada por:

$$\frac{\partial \eta_j}{\partial p_j} = \frac{\sum_{l=j+1}^r p_l}{p_j} \times \frac{\sum_{l=j+1}^r p_l}{\left(\sum_{l=j+1}^r p_l \right)^2} = \frac{1}{p_j}, \quad j = 1, 2, \dots, r-1$$

Dessa forma,

$$z_{ij} = \eta_{ij}^0 + (y_{ij} - p_{ij}^0) \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0 = \eta_j + \frac{(y_j - p_j)}{p_j}, \quad j=1,2,\dots,r-1$$

$$\text{E } w_{ij}^{-1} = \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0^2 V_{ij}^0 = \left(\frac{1}{p_j} \right)^2 p_j (1 - p_j) = \frac{(1 - p_j)}{p_j}, \text{ por tanto,}$$

$$w_i = \frac{1}{w_i^{-1}} = \frac{p_j}{(1 - p_j)}, \quad j=1,2,\dots,r-1.$$

Como este modelo também é cumulativo num certo sentido, também é necessário um processo sequencial para a obtenção dos p 's partindo-se de um valor inicial, que neste caso também será p_1 :

$$\eta_1 = \ln \left(\frac{p_1}{\sum_{m=2}^r p_m} \right), \text{ onde, substituindo-se (ii), tem-se igual resultado ao modelo}$$

cumulativo:

$$\exp(\eta_1) = \frac{p_1}{1 - p_1}, \text{ em que, isolando-se o } p_1, \text{ tem-se que } p_1 = \frac{\exp(\eta_1)}{1 + \exp(\eta_1)}.$$

$$\text{Como } \eta_2 = \ln \left(\frac{p_2}{\sum_{m=3}^r p_m} \right), \text{ de (iii) temos, ao isolar } p_2 \text{ e conhecendo-se } p_1 \text{ e } \eta_2$$

que

$$p_2 = \frac{\exp(\eta_2)}{1 + \exp(\eta_2)} \times (1 - p_1).$$

Os demais p 's são obtidos de maneira análoga pelo processo:

$$p_j = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} \times \left(1 - \sum_{m=1}^{j-1} p_m \right), \quad j=2,\dots,r-1.$$

$$\text{Para obtenção de } p_r, \text{ tendo os demais valores de } p, \text{ como } \eta_{r-1} = \ln \left(\frac{p_{r-1}}{p_r} \right),$$

que:

$$p_r = \frac{p_{r-1}}{\exp(\eta_{r-1})}.$$

Modelo de Categorias Adjacentes (ou adjacent-category logistic model):

Neste caso, considere o seguinte preditor linear:

$$\eta_j = \ln\left(\frac{p_j}{p_{j+1}}\right), j = 1, 2, \dots, r-1$$

onde p_j é a probabilidade de um determinado evento ter resposta j , dadas as covariáveis incluídas no modelo.

Daí, tem-se que a derivada de η_j em relação a p_j é dada por:

$$\frac{\partial \eta_j}{\partial p_j} = \frac{p_{j+1}}{p_j} \times \frac{p_{j+1}}{(p_{j+1})^2} = \frac{1}{p_j}, j = 1, 2, \dots, r-1$$

Dessa forma,

$$z_{ij} = \eta_{ij}^0 + (y_{ij} - p_{ij}^0) \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0 = \eta_j + \frac{(y_j - p_j)}{p_j}, j = 1, 2, \dots, r-1$$

$$E \quad w_{ij}^{-1} = \left(\frac{\partial \eta_{ij}}{\partial p_{ij}} \right)_0^2 V_{ij}^0 = \left(\frac{1}{p_j} \right)^2 p_j (1 - p_j) = \frac{(1 - p_j)}{p_j}, \text{ por tanto,}$$

$$w_i = \frac{1}{w_i^{-1}} = \frac{p_j}{(1 - p_j)}, j = 1, 2, \dots, r-1.$$

A obtenção dos valores de p neste caso é um pouco mais complicada que nos demais modelos, pois para cada valor considera-se todos os η estimados.

$$\text{Partindo da ideia de que } \exp(\eta_{r-1}) = \frac{p_{r-1}}{p_r};$$

$$\exp(\eta_{r-1}) \times \exp(\eta_{r-2}) = \frac{p_{r-1}}{p_r} \times \frac{p_{r-2}}{p_{r-1}} = \frac{p_{r-2}}{p_r};$$

$$\dots; \prod_{m=1}^{r-1} \exp(\eta_m) = \exp\left(\sum_{m=1}^{r-1} \eta_m\right) = \frac{p_1}{p_r}.$$

$$\text{Juntando isso ao fato de que se } \sum_{m=1}^{r-1} p_m = 1 - p_r, \text{ então } \frac{\sum_{m=1}^{r-1} p_m}{p_r} = \frac{1 - p_r}{p_r}, \text{ temos}$$

que:

$$\sum_{l=1}^{r-1} \exp\left(\sum_{m=l}^{r-1} \eta_m\right) = \sum_{m=1}^{r-1} \frac{p_m}{p_r} = \frac{\sum_{m=1}^{r-1} p_m}{p_r} = \frac{1-p_r}{p_r}, \text{ e daí, isolando-se o } p_r, \text{ tem-se:}$$

$$p_r = \frac{1}{1 + \sum_{l=1}^{r-1} \exp\left(\sum_{m=l}^{r-1} \eta_m\right)}.$$

Como $\exp\left(\sum_{m=j}^{r-1} \eta_m\right) = \frac{p_j}{p_r}$, $j = 1, 2, \dots, r-1$, obtem-se os demais valores de p da

seguinte maneira:

$$p_j = \exp\left(\sum_{m=j}^{r-1} \eta_m\right) \times p_r, \quad j = 1, 2, \dots, r-1.$$

10.2 Solução para obter o risco relativo espacial na análise cumulativa.

Seja r o número de categorias da variável reposta.

Considere $r=3$:

Seja $\varphi_{ij}(\mathbf{x}, \mathbf{u}) = \exp(\hat{\rho}_{ij}(\mathbf{x}, \mathbf{u}))$, da equação (9), temos que:

$$\varphi_1(\mathbf{x}, \mathbf{u}) = \frac{q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u})}, \text{ equivalente a}$$

$$q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) = \varphi_1(\mathbf{x}, \mathbf{u}) q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + \varphi_1(\mathbf{x}, \mathbf{u}) q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}), \text{ multiplicando os dois lados por } \varphi_2(\mathbf{x}, \mathbf{u}),$$

$$\varphi_2(\mathbf{x}, \mathbf{u}) q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) = \varphi_2(\mathbf{x}, \mathbf{u}) \varphi_1(\mathbf{x}, \mathbf{u}) q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + \varphi_2(\mathbf{x}, \mathbf{u}) \varphi_1(\mathbf{x}, \mathbf{u}) q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}) \quad (\text{i})$$

e

$$\varphi_2(\mathbf{x}, \mathbf{u}) = \frac{q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u})}, \text{ equivalente a}$$

$$q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) = \varphi_2(\mathbf{x}, \mathbf{u}) q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}), \text{ multiplicando os dois lados por } \varphi_1(\mathbf{x}, \mathbf{u}),$$

$$\varphi_1(\mathbf{x}, \mathbf{u}) q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + \varphi_1(\mathbf{x}, \mathbf{u}) q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) = \varphi_1(\mathbf{x}, \mathbf{u}) \varphi_2(\mathbf{x}, \mathbf{u}) q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}) \quad (\text{ii})$$

Subtraindo (ii) de (i), temos:

$[\varphi_2(\mathbf{x}, \mathbf{u}) - \varphi_1(\mathbf{x}, \mathbf{u})]q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) - \varphi_1(\mathbf{x}, \mathbf{u})q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u}) = \varphi_1(\mathbf{x}, \mathbf{u})\varphi_2(\mathbf{x}, \mathbf{u})q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u})$, igual a $\hat{\lambda}_1(\mathbf{x}, \mathbf{u})[\varphi_2(\mathbf{x}, \mathbf{u})q_1 - \varphi_1(\mathbf{x}, \mathbf{u})q_1] = \hat{\lambda}_2(\mathbf{x}, \mathbf{u})[\varphi_1(\mathbf{x}, \mathbf{u})\varphi_2(\mathbf{x}, \mathbf{u})q_2 + \varphi_1(\mathbf{x}, \mathbf{u})q_2]$, equivalente

a

$$\frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} = \frac{[\varphi_2(\mathbf{x}, \mathbf{u})q_1 - \varphi_1(\mathbf{x}, \mathbf{u})q_1]}{[\varphi_1(\mathbf{x}, \mathbf{z})\varphi_2(\mathbf{x}, \mathbf{u})q_2 + \varphi_1(\mathbf{x}, \mathbf{u})q_2]} \quad (\text{iii})$$

De $\varphi_2(\mathbf{x}, \mathbf{u})$, temos:

$$\varphi_2(\mathbf{x}, \mathbf{u}) = \frac{q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{q_3\hat{\lambda}_3(\mathbf{x}, \mathbf{u})} \stackrel{+\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{=} \frac{q_1 + q_2 \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{q_3 \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}, \text{ daí, utilizando (iii),}$$

$$\frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} q_3 \varphi_2(\mathbf{x}, \mathbf{u}) = q_1 + q_2 \frac{[\varphi_2(\mathbf{x}, \mathbf{u})q_1 - \varphi_1(\mathbf{x}, \mathbf{u})q_1]}{[\varphi_1(\mathbf{x}, \mathbf{u})\varphi_2(\mathbf{x}, \mathbf{u})q_2 + \varphi_1(\mathbf{x}, \mathbf{u})q_2]}, \text{ equivalente a}$$

$$\frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} = \frac{q_1}{q_3\varphi_2(\mathbf{x}, \mathbf{u})} + \frac{[\varphi_2(\mathbf{x}, \mathbf{u})q_1 - \varphi_1(\mathbf{x}, \mathbf{u})q_1]}{q_3\varphi_2(\mathbf{x}, \mathbf{u})[\varphi_1(\mathbf{x}, \mathbf{u})\varphi_2(\mathbf{x}, \mathbf{u}) + \varphi_1(\mathbf{x}, \mathbf{u})]} \quad (\text{iv}).$$

Para obter os riscos relativos espaciais estimados, a partir de (iii) e (iv):

$$\begin{aligned} \hat{\theta}_1(\mathbf{x}, \mathbf{u}) &= \frac{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + \hat{\lambda}_3(\mathbf{x}, \mathbf{u})} \stackrel{+\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{=} \frac{\frac{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{\frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} = \frac{1}{\frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} \text{ e} \\ \hat{\theta}_2(\mathbf{x}, \mathbf{u}) &= \frac{\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + \hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})} \stackrel{+\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{=} \frac{\frac{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{\frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} = \frac{1 + \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}}{\frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}} \end{aligned}$$

Considere agora $r = 4$

Seja $\varphi_i = \exp(\hat{\rho}_i(\mathbf{x}, \mathbf{u}))$ e $\hat{\lambda}_i = \hat{\lambda}_i(\mathbf{x}, \mathbf{u})$, da equação (9), temos que:

$$\varphi_1(\mathbf{x}, \mathbf{u}) = \frac{q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u})}{q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + q_3\hat{\lambda}_3(\mathbf{x}, \mathbf{u}) + q_4\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}; \quad \varphi_2(\mathbf{x}, \mathbf{u}) = \frac{q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{q_3\hat{\lambda}_3(\mathbf{x}, \mathbf{u}) + q_4\hat{\lambda}_4(\mathbf{x}, \mathbf{u})};$$

$$\varphi_3(\mathbf{x}, \mathbf{u}) = \frac{q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) + q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + q_3\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{q_4\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}$$

De maneira equivalente;

$$q_1\hat{\lambda}_1(\mathbf{x}, \mathbf{u}) = \varphi_1(\mathbf{x}, \mathbf{u})q_2\hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + \varphi_1(\mathbf{x}, \mathbf{u})q_3\hat{\lambda}_3(\mathbf{x}, \mathbf{u}) + \varphi_1(\mathbf{x}, \mathbf{u})q_4\hat{\lambda}_4(\mathbf{x}, \mathbf{u});$$

$$q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) = -q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) + \varphi_2 q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}) + \varphi_2 q_4 \hat{\lambda}_4(\mathbf{x}, \mathbf{u}) ;$$

$$q_1 \hat{\lambda}_1(\mathbf{x}, \mathbf{u}) = -q_2 \hat{\lambda}_2(\mathbf{x}, \mathbf{u}) - q_3 \hat{\lambda}_3(\mathbf{x}, \mathbf{u}) + \varphi_3 q_4 \hat{\lambda}_4(\mathbf{x}, \mathbf{u}) .$$

Que é igual a:

$$\begin{cases} 1 = \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_2}{q_1} \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_3}{q_1} \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \frac{\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \\ 1 = -\frac{q_2}{q_1} \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \varphi_2(\mathbf{x}, \mathbf{u}) \frac{q_3}{q_1} \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \varphi_2(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \frac{\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \\ 1 = -\frac{q_2}{q_1} \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} - \frac{q_3}{q_1} \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} + \varphi_3(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \frac{\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \end{cases} , \text{ daí temos três}$$

equações com três incógnitas, o que se resume ao sistema linear:

$$\begin{bmatrix} \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_2}{q_1} & \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_3}{q_1} & \varphi_1(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \\ -\frac{q_2}{q_1} & \varphi_2(\mathbf{x}, \mathbf{u}) \frac{q_3}{q_1} & \varphi_2(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \\ -\frac{q_2}{q_1} & -\frac{q_3}{q_1} & \varphi_3(\mathbf{x}, \mathbf{u}) \frac{q_4}{q_1} \end{bmatrix} \times \begin{bmatrix} \frac{\hat{\lambda}_2(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \\ \frac{\hat{\lambda}_3(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \\ \frac{\hat{\lambda}_4(\mathbf{x}, \mathbf{u})}{\hat{\lambda}_1(\mathbf{x}, \mathbf{u})} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} ,$$

Como as razões amostrais e os valores exponenciais dos logitos ajustados pelo modelo GAM são sempre positivos, garante-se que este sistema tem solução única.

Para valores de r acima de quatro pode-se seguir o mesmo raciocínio utilizado para $r=3$ e $r=4$.

10.3 Análises de resíduos dos acidentes de trabalho.

Com base nas medidas definidas na seção 4.7 sobre os resíduos de Pearson e resíduos *deviance* para GAM multinomiais, obtemos, para as análises dos acidentes de trabalho, os gráficos apresentados nas Figuras 6 a 9.

Figura 6: Análise dos resíduos do modelo binomial apenas com informação espacial dos acidentes de trabalho em Piracicaba, 2007.

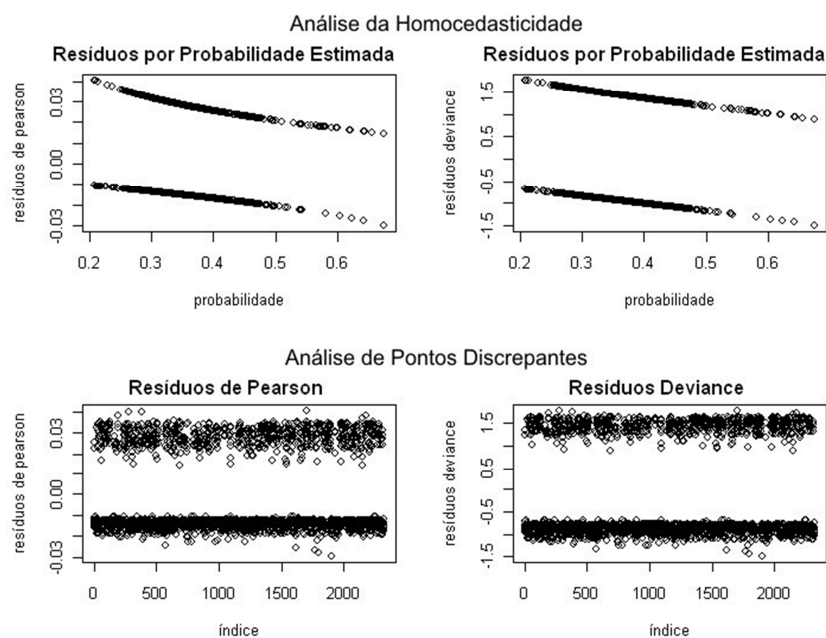


Figura 7: Análise dos resíduos do modelo binomial semiparamétrico dos acidentes de trabalho em Piracicaba, 2007.

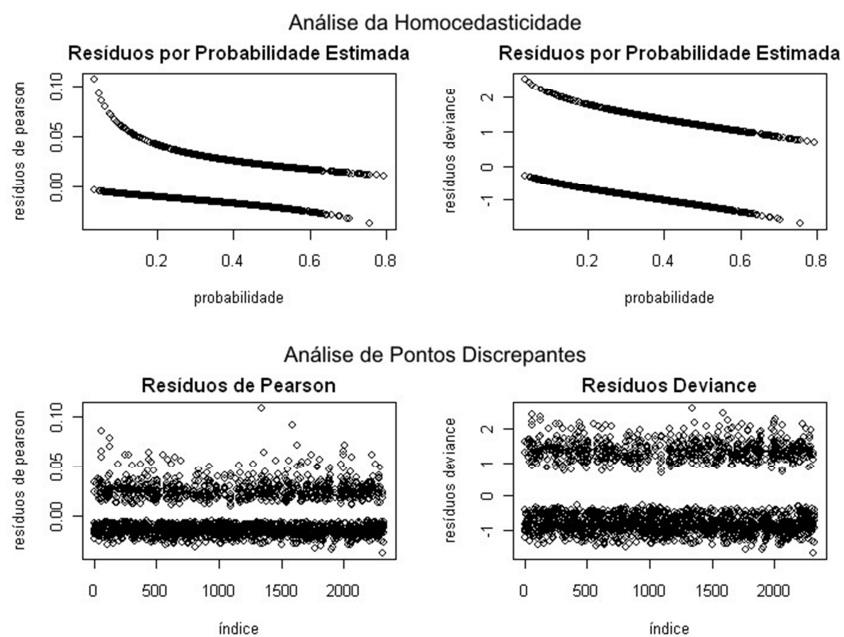
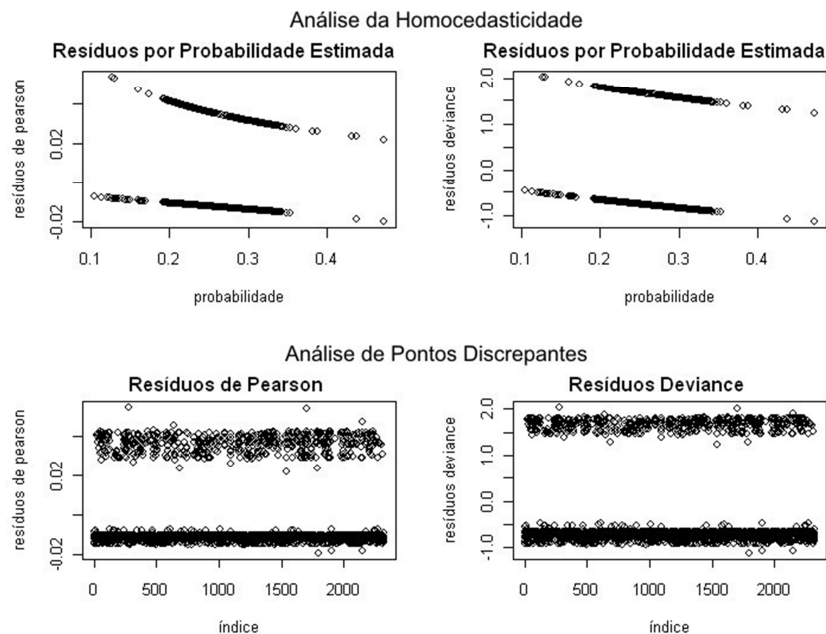


Figura 8: Análise dos resíduos do modelo multinomial apenas com informação espacial dos acidentes de trabalho em Piracicaba, 2007. A) Logito entre acidentes leves e controles. B) Logito entre acidentes graves e controles.

A)



B)

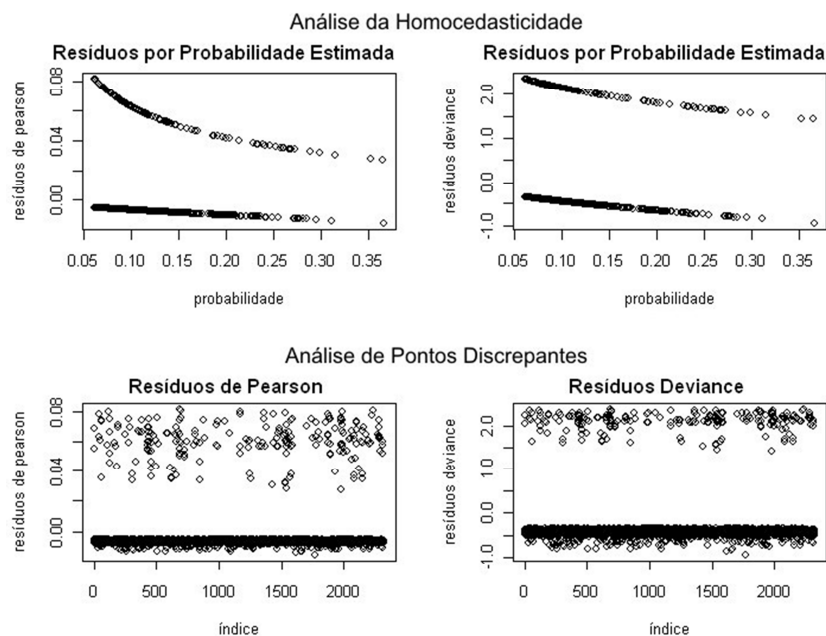
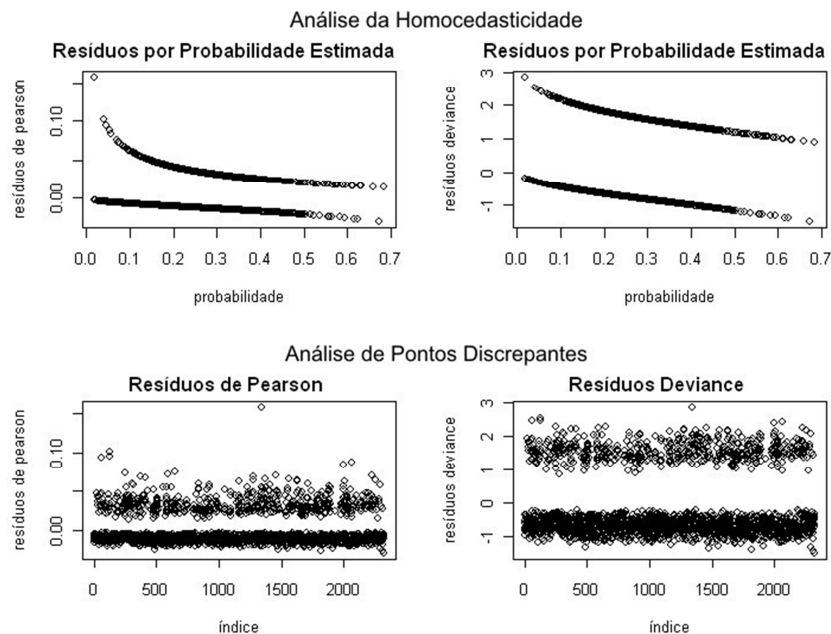
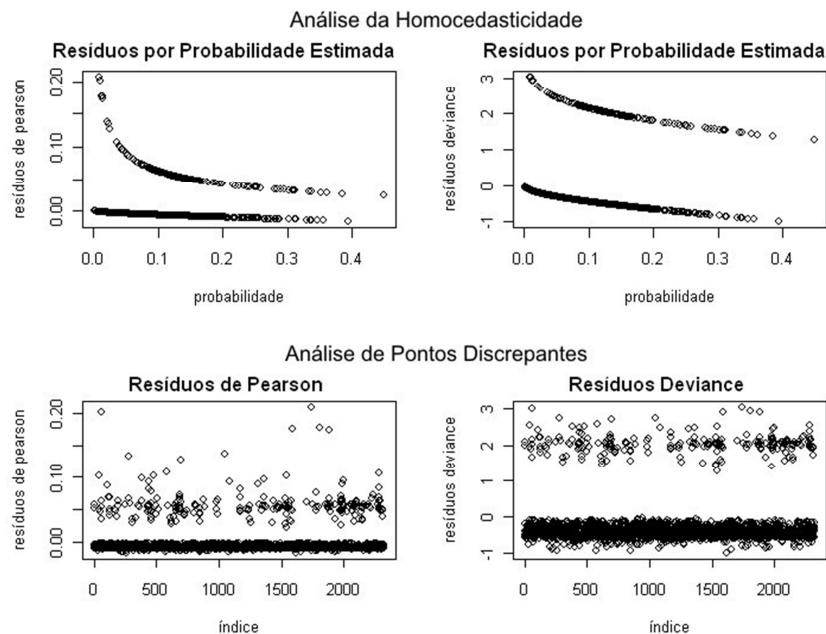


Figura 9: Análise dos resíduos do modelo multinomial semiparamétrico dos acidentes de trabalho em Piracicaba, 2007. A) Logito entre acidentes leves e controles. B) Logito entre acidentes graves e controles.

A)



B)



10.4 Análises de resíduos dos criadouros de mosquitos.

Para o estudo dos criadouros de mosquitos, os gráficos com as análises de resíduos dos GAM multinomiais estão nas Figuras 10 a 14.

Figura 10: Análise dos resíduos do modelo binomial semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007.

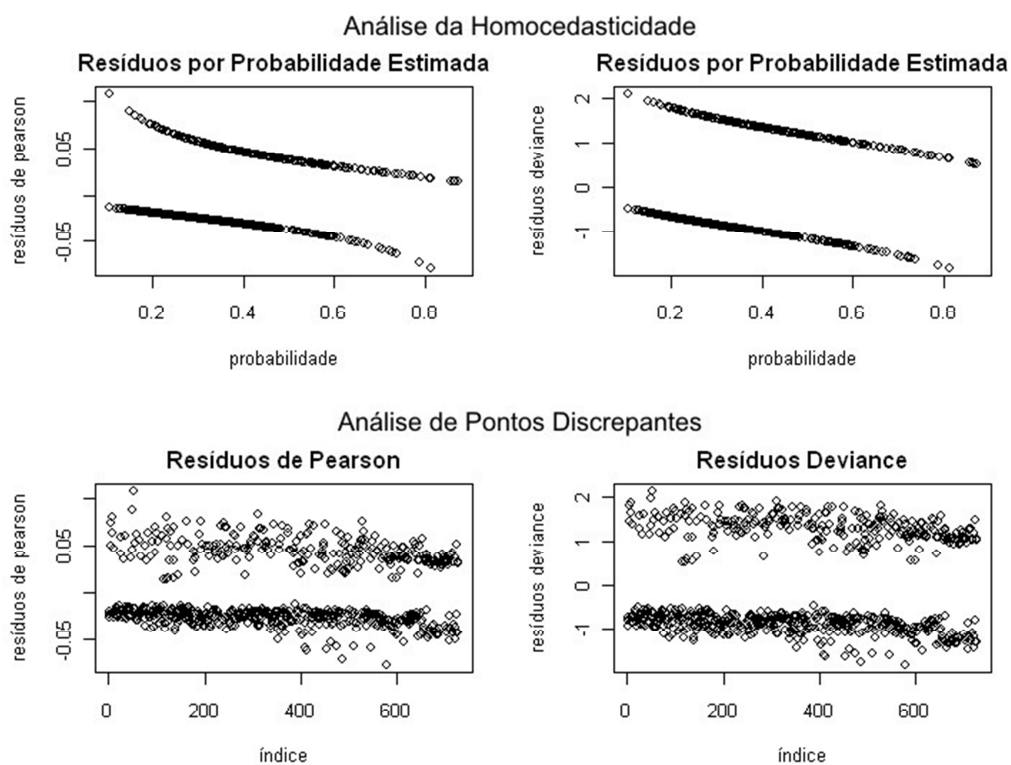
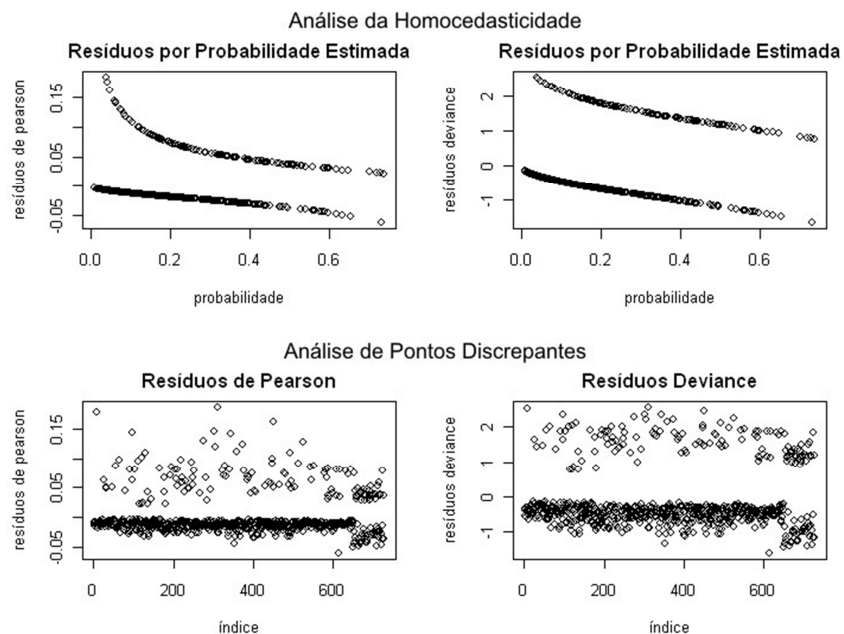


Figura 11: Análise dos resíduos do modelo cumulativo semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007.

A) $\text{logito}(P[Y \leq 1])$



B) $\text{logito}(P[Y \leq 2])$

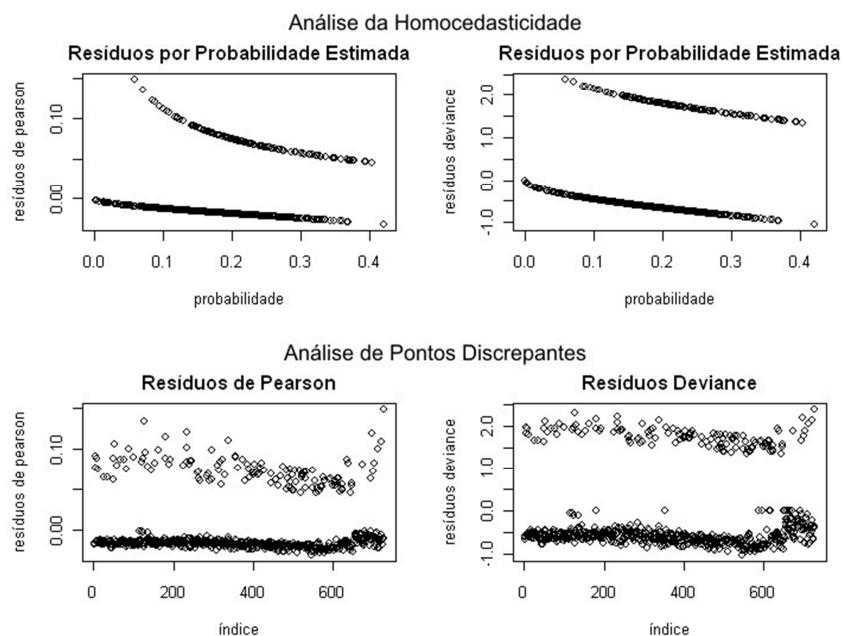
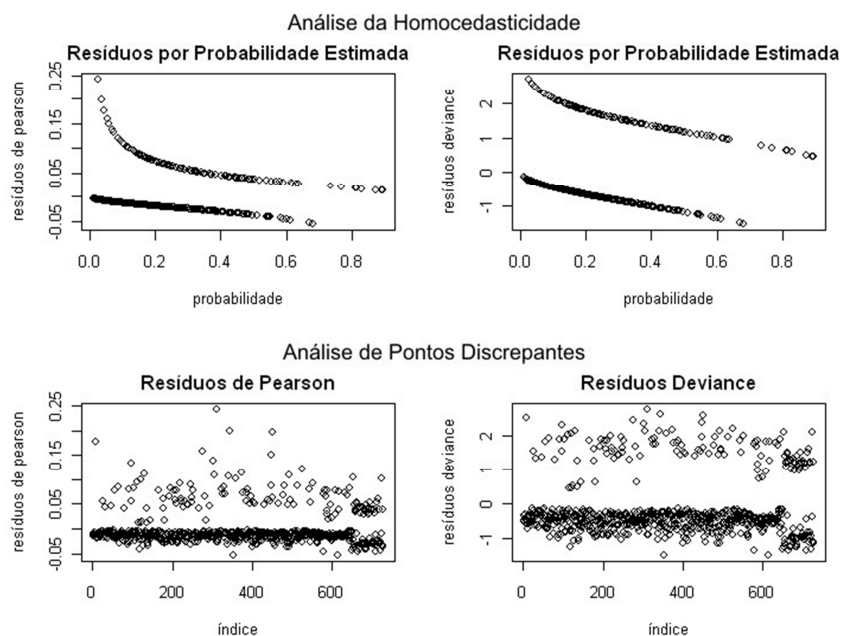


Figura 12: Análise dos resíduos do modelo de razões sequenciais semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007.

A) $\text{logito}(P[Y=1 | Y \geq 1])$



B) $\text{logito}(P[Y=2 | Y \geq 2])$

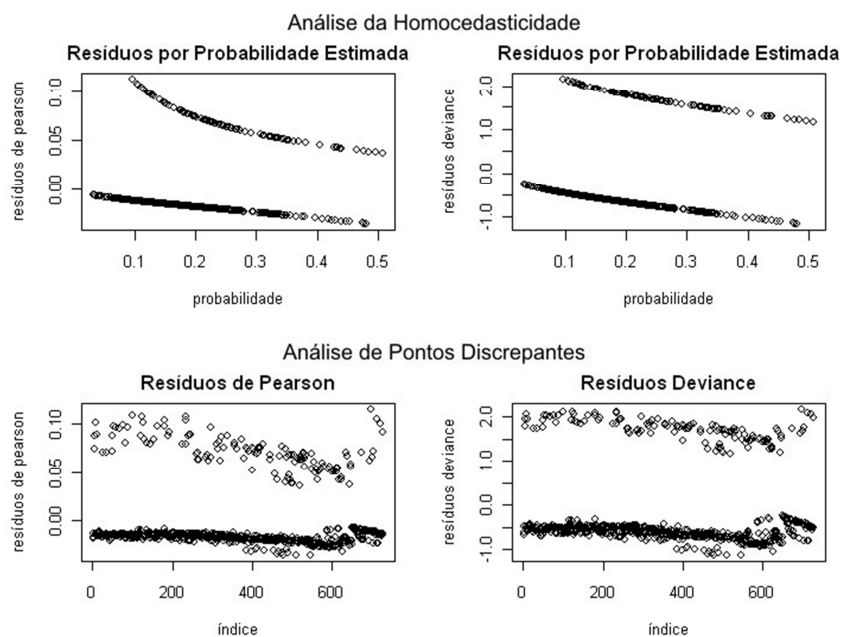
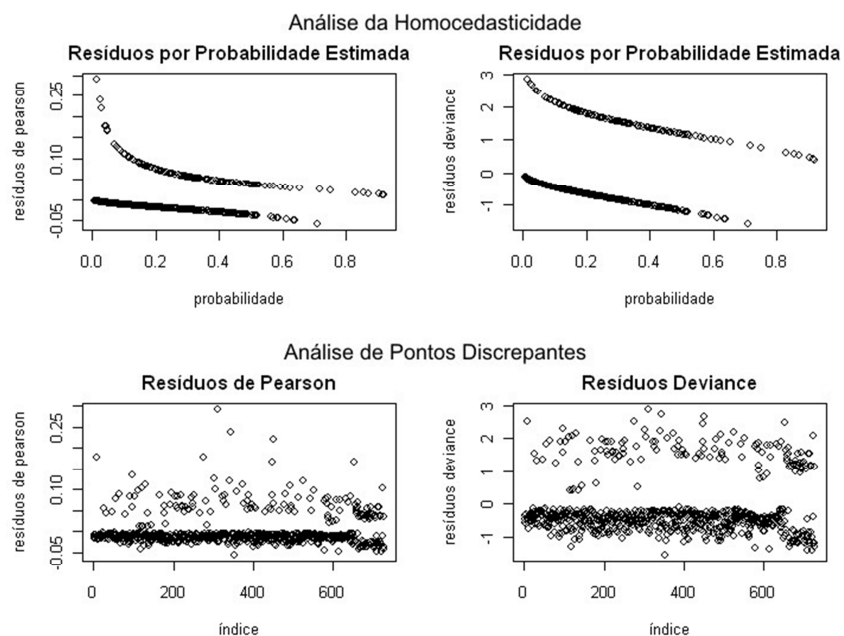


Figura 13: Análise dos resíduos do modelo de categorias adjacentes semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007.

A) $\log(P[Y=1]/P[Y=2])$



B) $\log(P[Y=2]/P[Y=3])$

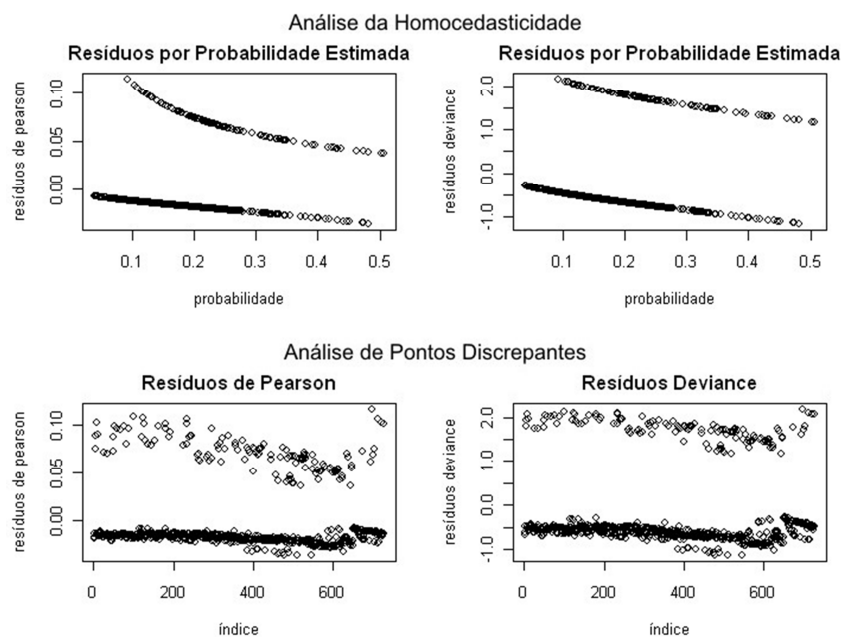
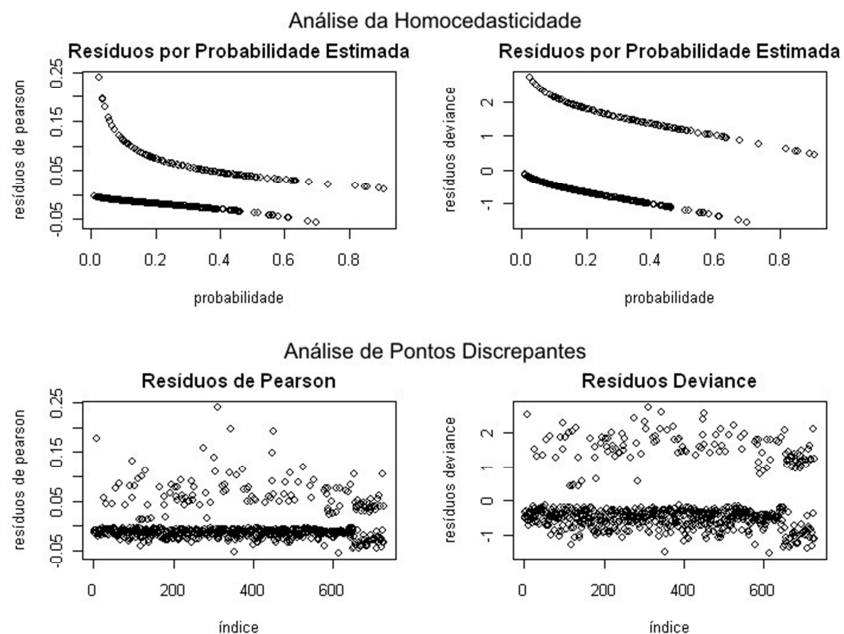
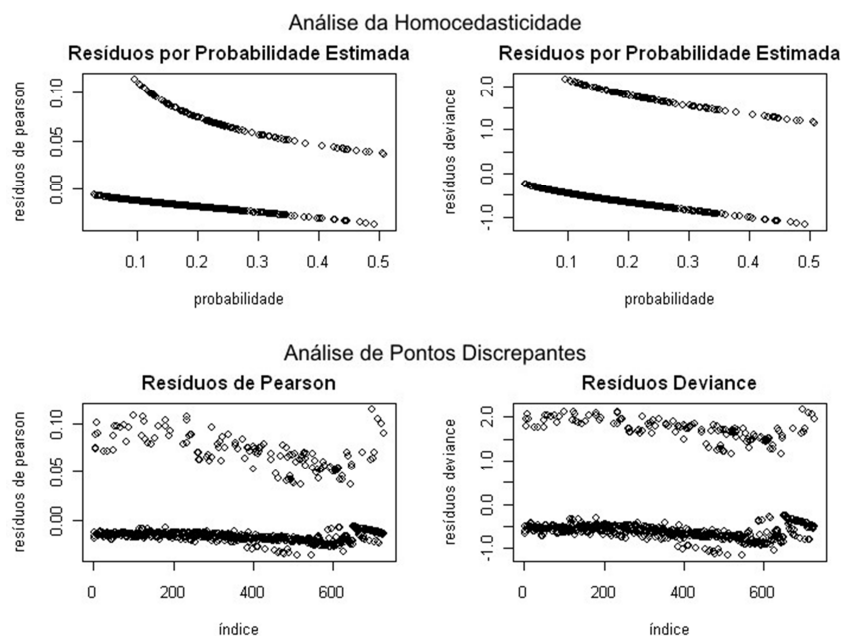


Figura 14: Análise dos resíduos do modelo politômico semiparamétrico de possíveis criadouros de mosquito no Distrito Sul de Campinas, 2007.

A) $\log(P[Y=1]/P[Y=3])$



B) $\log(P[Y=2]/P[Y=3])$



10.5 Referências de artigos relacionados a esta tese.

Estão listados aqui apenas artigos com participação da autora desta tese que, de alguma forma, estão relacionados ao conteúdo desta.

Mafra ACCN, Nucci LB, Cordeiro R, Stephan C. Estudo caso-controle com resposta multinomial: uma proposta de análise. *Cadernos de Saúde Pública* 2010; **26**(3):451-459.

Cordeiro R, Donalisio MR, Andrade VR, Mafra ACCN, Nucci LB, Brown JC, Stephan C. Spatial distribution of the risk of dengue fever in southeast Brazil, 2006-2007. *BMC Public Health* 2011; **11**:355.