Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Uma proposta de modelagem para o risco de sofrer acidente de trabalho em Piracicaba/SP em estudos caso-controle espacial

Marcelo Tavares de Lima

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2010

Marcelo Tavares de Lima
Bacharel em Estatística

Uma proposta de modelagem para o risco de sofrer acidente de trabalho em Piracicaba/SP em estudos caso-controle espacial

Orientador:
Prof . Dr. **PAULO JUSTINIANO RIBEIRO Jr.**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2010

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Lima, Marcelo Tavares de
Uma proposta de modelagem para o risco de sofrer acidente de trabalho em Piracicaba/SP em estudos caso-controle espacial. - - Piracicaba, 2010.
84 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura “Luiz de Queiroz”, 2010.
Bibliografia.

1. Acidente de trabalho - Modelagem - Piracicaba, SP 2. Distribuição espacial 3. Estatística computacional 4. Linguagem de programação 5. Modelos lineares generalizados 6. Riscos ocupacionais I. Título

CDD 519.50285
L732p

Dedicatória

a *Deus*, acima de tudo. Aos meus pais, familiares e amigos. A todos que fizeram e fazem parte da minha vida.

AGRADECIMENTOS

Desejo externar os meus agradecimentos aos meus pais Waldir Gomes de Lima e Raimunda Tavares de Araújo e a todos os meus irmãos Regina, Claudia, Rodrigo, Vanessa, Jennifer e Jéssica pois, apesar da distância me deram muito incentivo, valorizaram minhas decisões, ajudaram em muitos momentos, o que me deu a certeza de que estarão sempre ao meu lado.

Ao Prof. Dr. Paulo Justiniano Ribeiro Junior e sua equipe maravilhosa composta pelos professores Wagner Bonat, Gledson Picharski entre outros, pelo seu incentivo e esforço em me ajudar ao longo desta jornada. E, também ao Celso Stepahn que me ajudou muito nessa trajetória. Ao Prof. Dr. Ricardo Cordeiro, do Departamento de Medicina Preventiva da Unicamp, por ter me permitido utilizar seus dados e também pelo apoio recebido. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo auxílio financeiro prestado.

Aos professores do Departamento de Ciências Exatas da ESALQ/USP, Prof.Dra. Clarice Demétrio, Prof.Dr. Carlos Tadeu, Prof.Dra. Roseli Leandro, Prof.Dra. Sônia Piedade, Prof.Dr. César Gonçalves e todos os outros professores com os quais não tive um contato em sala de aula mas que convivi no dia-a-dia. Aos funcionários Solange Sabadin, Luciane Brajão, Jorge, Eduardo e todos os outros, os quais mostraram ser altamente profissionais e humanos ao lidar conosco.

À Dra. Lilia T. Montali, pesquisadora do NEPP/Unicamp pois sem ela não teria conseguido chegar aonde cheguei, assim como, a Dra. Stella Barberá, que me ajudou tanto quanto e todos os demais amigos e colegas de trabalho do NEPP/UNICAMP e do NEPO/UNICAMP.

Ao Alexandre Barbosa (*in memorian*) por tê-lo tido como amigo por pouco tempo mas, que foi um quase irmão. Aos demais colegas de turma, Rodrigo, Epaminondas, Tiago, Diógenes, Keliny e Renato, pelo companheirismo. Ao Ronaldo, Márcio e Everton por terem tido paciência nos meus momentos de desespero e pela companhia nos momentos de alegria. Aos colegas e amigos da pós-graduação que me ajudaram, de diferentes formas, em especial à Mariana Urbano que me deu dicas importantíssimas em todos os momentos. Ao Gledson Picharski, Mariana, Suelen e Edicléia, alunos do curso de graduação em estatística da

UFPR pela colaboração prestada.

Aos professores do Departamento de Estatística da UFAM por terem contribuído para minha formação de forma significativa, em especial Prof.Dra. Maria Ivanilde, Prof. Dra. Rosana Parente e Prof. Dr. Celso Rômulo.

Aos meus amigos de Piracicaba, Campinas e Manaus e, em especial a minha amiga Marli Soares que nos meus momentos mais difíceis teve paciência em me escutar a desabafar.

SUMÁRIO

RESUMO 9
ABSTRACT 11
LISTA DE FIGURAS 13
LISTA DE TABELAS 15
1 INTRODUÇÃO 17
2 DESENVOLVIMENTO 23
2.1 Análise espacial 23
2.2 Estatística espacial 24
2.3 Análise de processos pontuais 24
2.4 Análise de dados de área 25
2.5 Modelos lineares generalizados 26
2.6 Modelos aditivos generalizados 27
2.7 Estudos caso-controle espaciais 28
2.8 Modelos estruturados aditivamente 31
3 MATERIAL E MÉTODOS 39
3.1 Material 39
3.2 Métodos 40
4 RESULTADOS E DISCUSSÃO 45
4.1 Análise descritiva 45
4.2 Análise inferencial 50
4.2.1 Abordagem com modelos estruturados aditivamente 55
5 CONSIDERAÇÕES FINAIS 65
REFERÊNCIAS 67
APÊNDICE 71

RESUMO

Uma proposta de modelagem para o risco de sofrer acidente de trabalho em Piracicaba/SP em estudos caso-controle espaciais

O mapeamento e a estimação de riscos e incidências são ferramentas muito úteis para a Epidemiologia pois, auxiliam na prevenção de agravos da saúde e, também auxiliam no planejamento e avaliação dos serviços de saúde. Este trabalho busca utilizar uma ferramenta estatística que incorpora de forma adequada este tipo de análise ao estudo de outras características que estejam relacionadas a estes agravos. No presente trabalho utiliza-se como aplicação dados do estudo caso-controle espacial com base populacional de acidentes de trabalho com a proposta de estimar a distribuição espacial do risco de sofrer acidente de trabalho na área urbana do município de Piracicaba/SP entre trabalhadores que se encontravam na situação de precarização do trabalho em associação com outras variáveis de interesse através de modelos aditivos generalizados (MAG) e, através disso, mostrar que ao incorporar de forma explícita o espaço no processo de modelagem dos dados ocorre um ganho significativo na explicação da variação do risco. O modelo MAG utilizado tem variável resposta binomial (caso e controle) e multinomial (caso e controle separados pela gravidade do acidente sofrido). Com os modelos ajustados, mapas foram desenhados com indicações de diferentes cores para a intensidade do risco de sofrer acidente de trabalho. Outra abordagem utilizada para os dados espaciais de acidentes de trabalho foi a INLA (INTEGRATED NESTED LAPLACE APPROXIMATIONS), a qual é utilizada como processo de modelagem para a família dos modelos Gaussianos latentes através de novos métodos para esta família de modelos. A intenção foi mostrar como essa nova abordagem lida com dados do tipo espacial e, fazer uma comparação com a abordagem feita pela modelagem GAM.

Palavras-chaves: Caso-controle; Estatística espacial; Modelos lineares generalizados; Modelos aditivos generalizados; Modelos Gaussianos latentes; INLA; Software R

ABSTRACT

One approach model for the risk of accidents at work in Piracicaba-SP in case-control space studies

Mapping and estimation of risks and impacts are very useful tools for Epidemiology at the assistance in prevention of injuries and health, also assists in planning and evaluation of health services. This paper seeks to use a statistical tool that adequately incorporates this type of analysis to the study of other characteristics that are related these illnesses. In the present work is used as application data from case-control study space-based population accidents with the proposal to estimate the spatial distribution of risk of suffering an accident at work in the urban area of Piracicaba/SP among workers who were in employed as casual labor in combination with other variables of interest using generalized additive models (GAM) and, thereby, show that by incorporating explicitly space in the process of data modeling is a gain significant in explaining the variation in risk. The GAM model have used binomial response variable (case and control) and multinomial (case and control separated by the severity of the accident suffered). With the adjusted models, maps were drawn with indications of different colors to the intensity of the risk of accident. Another approach used for spatial data on accidents at work was the INLA (INTEGRATED NESTED LAPLACE APPROXIMATIONS), which is used as a modeling process for the family of latent Gaussian models through new methods for this family of models. The intention was to show how this new approach deals with spatial data and a comparison with the approach made by GAM modeling.

Keywords: Case-control study; Spatial statistics; Generalized linear models; Generalized additive models; latent Gaussian models; INLA; Software R

LISTA DE FIGURAS

Figura 1 - Mapa de casos e controles de acidente de trabalho, segundo setores censitários 40
Figura 2 - Esquema do estimador de *kernel* 42
Figura 3 - Mapa de *kernel* 49
Figura 4 - Função K cruzada com envelopes de simulação via MCMC 50
Figura 5 - Mapa de kernel após o deslocamento dos pontos coincidentes 51
Figura 6 - Função K cruzada com envelopes de simulação após o deslocamento dos pontos coincidentes 52
Figura 7 - Número de casos, número de controles por setor censitário 56
Figura 8 - Razão entre o número de casos e o número de controles por setor censitário 57
Figura 9 - Efeito espacial estruturado 58
Figura 10 - Histograma do número de casos 59
Figura 11 - Distribuição a posteriori - Modelo com efeito espacial estruturado 60
Figura 12 - Efeito espacial não estruturado 60
Figura 13 - Modelo com efeito espacial estruturado para a idade média 62
Figura 14 - Modelo com efeito espacial estruturado para a proporção de pessoas com carteira assinada 63
Figura 15 - Modelo com efeito espacial estruturado para a proporção de trabalhadores domésticos 63

LISTA DE TABELAS

Tabela 1 - Distribuição do sexo segundo casos e controles - Piracicaba - 2006/2007 . . 45
Tabela 2 - Distribuição etária segundo casos e controles - Piracicaba - 2006/2007 . . . 46
Tabela 3 - Distribuição da escolaridade segundo casos e controles - Piracicaba - 2006/2007 47
Tabela 4 - Distribuição do risco referido segundo casos e controles - Piracicaba - 2006/2007 48
Tabela 5 - Distribuição da classificação da gravidade do acidente - Piracicaba - 2006/2007 48
Tabela 6 - Valores estimados para o MLG inicial 51
Tabela 7 - Valores estimados para o MLG obtido com o critério de seleção de variável por AIC 53
Tabela 8 - Valores estimados para o MLG com componente espacial suavizada 53
Tabela 9 - Valores estimados para o MLG com componente espacial suavizada para os dados com pontos coincidentes deslocados 54
Tabela 10 - Valores estimados para o MLG com componente espacial suavizada para os dados com pontos coincidentes deslocados 54
Tabela 11 - Estimativas de graus de liberdade efetivos (edf) para as funções semi-paramétricas relacionadas ao espaço e de odds ratios para as covariáveis paramétricas nos modelos ajustados para acidentes de trabalho em Piracicaba 55
Tabela 12 - Medidas resumo para a razão entre o número de casos e o número de controles segundo os setores censitários 56
Tabela 13 - Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal 57
Tabela 14 - Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal considerando a superdispersão nos dados 58
Tabela 15 - Ajuste do modelo com todas as covariáveis considerando efeito espacial estruturado e não estruturado 61

1 INTRODUÇÃO

Na tentativa de entender as origens de uma ocorrência não usual de um determinado evento, uma epidemia, por exemplo, os profissionais da saúde comumente buscam examinar a história individual de cada um dos acometidos procurando exposições compartilhadas por eles que não sejam freqüentes na população que deu origem a esses casos, na busca do entendimento dos mecanismos que governam o fenômeno em questão e, do efeito das desigualdades sobre a qualidade de vida, consequentemente, nas condições de saúde da população. O passo seguinte a ser dado por tais profissionais seria verificar se há relação entre o surgimento dos casos e as exposições eventualmente identificadas. Este é o campo da epidemiologia, área do conhecimento que estuda relações entre exposições e fenômenos do processo saúde/doença que ocorrem em populações humanas.

O entendimento de relações causais que poderiam explicar parcial ou totalmente o surgimento de séries de casos, como o acima exemplificado, é obtido comumente por meio da execução de estudos caso-controle. O desenvolvimento desses estudos constitui a maior contribuição metodológica da epidemiologia.

Os casos neste tipo de estudo surgem de uma população nem sempre facilmente identificável, chamada população fonte. A seleção dos controles, portanto, torna-se essencial para garantir a eficiência deste método e na medida em que o desenvolvimento da microinformática torna viável o desenvolvimento e a popularização do uso de sistemas de informação geográfica (SIG) e de ferramentas de análise espacial de dados, a dimensão espacial começa a ser incorporada ao método caso-controle.

Sabe-se que em muitos lugares a chance de ocorrer algum evento é maior do que em outros e, que a incidência de um determinado evento pode ser função do espaço geográfico. Nos anos 90 começou-se a desenvolver uma especialização dos estudos caso-controle que incorporam explicitamente o espaço ao conjunto de covariáveis que modelam o risco. São os chamados estudos caso-controle espaciais, que então passam a contribuir com a incorporação de métodos de análise espacial de dados.

Uma das formas de análise para este tipo de estudo é a análise de padrão espacial de pontos ou áreas, a qual pode ser definida como um conjunto de localizações numa determinada região de estudo representando o registro de eventos de interesse. Tais localizações

podem estar acompanhadas de informações adicionais relativas aos eventos registrados.

De um ponto de vista estatístico, um padrão observado de pontos pode ser modelado como a realização de um processo estocástico. O modelo mais simples de uma distribuição de pontos é o da aleatoriedade espacial completa (CSR[1]), no qual os eventos se distribuem de forma independente entre si sobre uma região de interesse. Esse modelo tem pouca aplicação epidemiológica (exceção feita a estudos em escalas grandes) uma vez que as populações fontes de casos distribuem-se elas próprias em aglomerados espaciais determinados por fatores ambientais e sociais. Na ausência de associação entre espaço e doença (hipótese nula), a ocorrência de casos espelha a heterogeneidade espacial da população fonte. A incorporação dos métodos de análise espacial de dados ao instrumental epidemiológico se deu a partir da indagação sobre o quanto um aglomerado observado de casos se deve à aglomeração de base da população fonte.

Para incorporar a dimensão espacial nos estudos é necessário conhecer bem o problema em questão, os métodos necessários, ter um conhecimento mínimo de sistemas de informação geográfica (SIG) e técnicas estatísticas apropriadas. Isso porque a existência de padrões espaciais implica a incorporação aos modelos estatísticos de estruturas de correlação entre as observações (CARVALHO; SOUZA-SANTOS, 2005).

A análise espacial de dados inclui um campo de pesquisas que teve grande desenvolvimento nos últimos anos, conhecido como estatística espacial. A estatística espacial é um ramo da estatística que estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possam ser indexados em relação a sua posição e modelados como realizações de processos estocásticos. Em estatística espacial a localização espacial do fenômeno estudado é utilizada de forma explícita em seu compreendimento.

O termo estatística espacial é usado para descrever uma ampla área de estudo de modelos e métodos para analisar dados espacialmente referenciados (DIGGLE; RIBEIRO, 2007). Ela é usualmente subdividida em três grandes áreas: geoestatística, dados de área e processos pontuais, as quais são utilizadas conforme o tipo de estudo realizado.

Uma das estruturas de modelagem adotada neste trabalho baseia-se em um processo pontual espacial, no qual se define uma medida de risco que varia continuamente sobre a região de estudo e estimada por meio de métodos semiparamétricos ou seja, aqui, por

[1]complete spatial randomness.

modelos aditivos generalizados (MAG). Essa abordagem possui a vantagem de permitir a incorporação no modelo de efeitos de determinantes individuais e ecológicos[2] de risco sob forma simples e de fácil interpretação. Também permite a construção de contornos de tolerância que auxiliam na identificação de áreas de alto/baixo risco e de um teste global da hipótese nula de risco constante relativa à região estudada. (SHIMAKURA et al., 2001).

Variáveis do tipo categóricas (dicotômicas ou polinomiais) tais como a presença ou ausência de determinados atributos nos indivíduos não são bem descritas por modelos lineares clássicos. Para estas variáveis, os modelos lineares generalizados ou não lineares podem ser mais apropriados. Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas, comumente estudadas separadamente, podem ser formuladas, de uma maneira unificada, como uma classe de modelos de regressão. A essa teoria de modelagem estatística, uma extensão dos modelos clássicos de regressão, foi dado o nome de modelos lineares generalizados (MLG).

Os modelos lineares generalizados incluem os modelos de regressão linear, modelos de análise de variância, modelos logit e probit para respostas dicotômicas, modelos log-lineares e de resposta multinomial para dados de contagem e outros modelos usados em dados de análise de sobrevivência. Eles permitem estudar padrões de variação sistemática da mesma forma que os modelos lineares clássicos são usados para estudar efeitos conjuntos entre tratamentos e covariáveis e são utilizados quando a distribuição de probabilidade do erro do modelo proposto não é normal.

O modelo de regressão linear generalizado logístico, o qual será tratado aqui com mais detalhes, é utilizado quando a variável resposta é qualitativa com dois ou mais resultados possíveis (categorias ou valores), o qual permite a predição de valores a partir de uma série de variáveis explicativas discretas e/ou contínuas. Ele é útil para modelar a probabilidade de um evento ocorrer como função de outros fatores.

As categorias ou valores que a variável resposta assume podem ser de natureza nominal ou ordinal. Em caso de natureza ordinal, há uma ordem natural entre as possíveis categorias e, então tem-se o contexto de regressão logística ordinal. Quando esta ordem não existe entre as categorias da(s) variável(is) independente(s) assume-se o contexto de regressão

[2]associado ao lugar.

logística nominal (FIGUEIRA, 2006). Outro contexto que a regressão logística pode assumir é a regressão logística multinomial, a qual é uma generalização da regressão logística usual. Ela é adequada quando a variável resposta é categórica e politômica, podendo assumir diversas categorias ou valores, mutuamente exclusivos e que não possuem qualquer ordenamento implícito. Portanto, a variável resposta assume distribuição multinomial.

Os dados deste estudo foram obtidos através de entrevistas realizadas em pronto-socorros da cidade de Piracicaba/SP e o intuito era coletar informações sobre trabalhadores, em situação de precariedade de trabalho[3], da área urbana do município e que sofreram algum tipo de acidente de trabalho.

Os modelos aditivos generalizados, uma extensão dos MLG surgem como uma opção prática, consistente e robusta para análise de doenças no espaço com funções semi-paramétricas, e, ao mesmo tempo, incorpora as análises de risco das outras exposições de interesse da forma comumente usada em modelos de regressão usuais.

Em algumas aplicações não é difícil encontrar situações onde os modelos usuais (MLG) deixam de ser adequados, por diversos fatores, tais como, inadequação de efeito estritamente linear no preditor, observações correlacionadas no espaço, observações correlacionadas no tempo, falta de interações complexas para modelar o efeito conjunto de algumas covariáveis, dentre outras.

Outra abordagem utilizada nestas situações é a utilização de modelos de regressão estruturados aditivamente, devido também, à flexibilidade existente nesta classe de modelos (FARHMEIR; TUTZ, 2001), cuja utilização em dados espaciais tem crescido amplamente. Esta, será utilizada na análise espacial de áreas, as quais serão os setores censitários da área urbana de Piracicaba totalizando 507 setores censitários.

A intenção deste trabalho é mostrar que ao incorporar de forma explícita o espaço na modelagem dos dados há um ganho significativo na explicação da variação do risco de sofrer acidente de trabalho considerando algumas informações (covariáveis) relacionadas a cada indivíduo entrevistado.

Outra intenção foi a de comparar as duas abordagens utilizadas (GAM e INLA),

[3]trabalhadores que não tinham a carteira de trabalho assinada ou, eram terceirizados, domésticos ou trabalhadores de rua.

a primeira considerando dados pontuais e a segunda, dados de área para o mesmo problema, cujo principal propósito foi o de mostrar o novo método de aproximação para a família de modelos Gaussianos latentes.

2 DESENVOLVIMENTO

2.1 Análise espacial

Dados espaciais são distinguidos por observações que são obtidas em localizações espaciais $s_1, s_2, \ldots, s_n$ onde s_i são coordenadas no plano $\mathbb{R}^2$ ou espaço $\mathbb{R}^3$.

A análise de dados espaciais pode ser realizada sempre que as informações estiverem espacialmente localizadas e quando for preciso levar em conta, explicitamente, a importância do arranjo espacial dos fenômenos na análise ou na interpretação dos resultados. Seu objetivo é aprofundar a compreensão do processo, avaliar evidências de hipóteses a ela relacionadas, ou ainda tentar prever valores em áreas onde as observações não estão disponíveis (BAILEY; GATTREL, 1995).

Os modelos de séries temporais, por exemplo, tentam modelar as correlações entres observações em diferentes tempos. Semelhantemente, com dados espaciais, a estrutura de correlação espacial necessita ser incorporada e modelada.

Segundo Bailey e Gattrel (1995), a análise espacial pode ser distinguida entre os vários métodos conforme a seguir:

- métodos essencialmente voltados para a visualização de dados espaciais;

- métodos exploratórios para investigar e resumir relações e padrões mapeados;

- métodos para especificação de modelos estatísticos e para estimar parâmetros.

A visualização gráfica é uma etapa fundamental da análise espacial. Através dela é possível identificar padrões espaciais nos dados, gerando hipóteses testáveis, bem como avaliar o ajuste de modelos propostos, ou ainda, a validade das previsões resultantes e, com isso, uma série de questões podem ser feitas, as quais podem ser respondidas com o auxílio de métodos gráficos ou estatísticas descritivas. Tais técnicas são conhecidas como Análise exploratória de dados espaciais, podendo ser classificadas em univariadas ou multivariadas, dependendo do número de variáveis envolvidas.

Dentre as técnicas univariadas, destacam-se os histogramas, mapas, estimativas de densidade, desenhos esquemáticos (boxplots) etc.; enquanto, entre as técnicas multivariadas

destacam-se as matrizes de dispersão, gráficos linked plots, gráficos de coordenadas paralelas entre outros.

2.2 Estatística espacial

A compreensão da distribuição espacial dos dados originários de fenômenos ocorridos no espaço é relevante para responder questões em diversas áreas do conhecimento. Diante desse desafio, vários métodos de análise estatística espacial foram desenvolvidos.

Estatística espacial é um nome genérico dado para o conjunto de métodos nos quais a localização dos dados (geometria) é relevante para a análise, ou seja, o conjunto de métodos estatísticos para análise exploratória e inferencial de dados espaciais é chamado de estatística espacial. Tais métodos são empregados quando se tem processos espaciais discretos, contínuos ou pontuais.

Bailey e Gatrell (1995) sugerem que a estatística espacial divide-se em quatro grandes áreas, de acordo com o tipo de dado analisado: análise de processos pontuais, análise de dados espacialmente contínuos (geoestatística), análise de dados de área e análise de dados de interação, enquanto outros sugerem que se divide em três áreas como dito anteriormente.

Para a análise espacial conceitos como dependência espacial e autocorrelação espacial são fundamentais. Entende-se por dependência espacial o fato de que a maior parte das ocorrências naturais ou sociais apresentam entre si uma relação que depende da distância (DRUCK et al., 2004). A expressão computacional do conceito de dependência espacial é a autocorrelação espacial. A idéia é verificar como a dependência espacial varia, a partir da comparação entre os valores de uma amostra e de seus vizinhos.

2.3 Análise de processos pontuais

A análise de pontos (ou eventos) tem como objetivo estudar a distribuição espacial de fenômenos que são expressos através de ocorrências identificadas como pontos localizados no espaço, também chamados de processos pontuais (ASSUNÇÃO, 2001).

Um padrão espacial de pontos é um conjunto de dados espaciais observados dentro de uma região A, compostos por coordenadas dos eventos de interesse. O objeto de interesse na análise destes tipos de dados é a própria localização espacial dos eventos em estudo.

É possível, também, atribuir covariáveis a essas localizações dos eventos.

Pode-se testar hipóteses sobre o padrão observado, como por exemplo, testar se o padrão é aleatório ou, apresentado em aglomerados ou pontos regularmente distribuídos.

Em geral, os pontos não estão associados a valores, e sim à ocorrência dos eventos considerados. Entretanto, em alguns casos, os pontos podem estar associados a atributos de identificação. Outra característica é que a área dos eventos não é considerada como sendo uma medida válida nos dados de distribuições pontuais. Por exemplo, a localização de acidentes, localizações de árvores em uma área de terra e a ocorrência de doenças são exemplos destes processos.

2.4 Análise de dados de área

A análise de áreas é aplicada quando não se dispõe da localização exata dos eventos, mas de um valor por área, isto é, é uma metodologia que lida com dados associados a levantamentos populacionais como censos e estatísticas de saúde, por exemplo. Estes dados se referem a indivíduos localizados em pontos específicos do espaço e são, em geral, agregados em unidades de análise delimitadas por polígonos fechados. Este tipo de análise usa atributos que não variam continuamente, mas que possuem valores específicos para subáreas (bairro, distrito, setor censitário, município etc.) que compõem uma dada região em estudo. O interesse está na detecção e possíveis explicações para o padrão espacial ou tendência de distribuição para valores de área.

Os tipos de análise de área incluem:

- Análise exploratória, com produção de indicadores de autocorrelação espacial, cujos métodos envolvem a procura de boas descrições dos dados, a fim de ajudar ao analista a desenvolver hipóteses sobre o assunto e modelos apropriados para tais dados (BAILEY; GATRELL, 1995).

- Modelos de regressão espacial, que buscam estabelecer relações entre duas ou mais variáveis, de modo que uma variável possa ser prevista a partir de outra, no caso de modelos univariados, ou a partir de outras, no caso de modelos multivariados.

- Modelos inferenciais, que buscam estabelecer a distribuição estatística subjacente aos da-

dos coletados. Este modelos incluem modelos bayesianos empíricos e completos (Câmara et al., 2003).

2.5 Modelos lineares generalizados

Por muito tempo os modelos lineares foram utilizados para descrever grande parte dos fenômenos aleatórios, mesmo quando este sob estudo não apresentava variável resposta com distribuição normal. Situação essa que era contornada com algum tipo de transformação.

Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLG) para resolver problemas do tipo citado acima. Essa proposta pode ser interpretada como uma generalização do modelo tradicional de regressão linear (PAULA, 2004).

Os MLG têm como idéia básica abrir o leque de opções para a distribuição da variável resposta, permitindo que esta pertença à família exponencial de distribuições, assim como, dar uma flexibilidade para a relação funcional entre o valor esperado da variável resposta e o preditor linear η, o qual é uma função linear nos parâmetros das variáveis explicativas do modelo.

Nelder e Wedderburn (1972) também propuseram um processo iterativo para a estimação dos parâmetros e introduziram o conceito de desvio (deviance) que tem sido largamente utilizado na avaliação da qualidade do ajuste dos MLG, bem como no desenvolvimento de resíduos e medidas de diagnóstico (PAULA, 2004).

Enfim, toda a estrutura conhecida para a regressão linear, pode ser estendida para os MLG. A grande vantagem é a possibilidade de se estudar conjuntamente as propriedades de diferentes modelos de regressão. No entanto, cada modelo tem propriedades próprias, as quais devem ser estudadas separadamente.

Um modelo linear generalizado é definido por: (1) variável resposta ou componente aleatório associado à distribuição da variável resposta, a qual tem distribuição pertencente à família exponencial; (2) um componente sistemático linear nos parâmetros, denominado preditor linear ou estrutura linear do modelo e; (3) uma função de ligação, a qual combina o componente aleatório e o componente sistemático (RESENDE; BIELE, 2002).

Supondo que $Y_1, \ldots, Y_n$ sejam variáveis aleatórias independentes, cada uma com

densidade dada na forma

$$f(y;\theta,\phi) = \exp\{\phi^{-1}[y\theta - b(\theta)] + c(y,\phi)\} \tag{1}$$

onde $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, $\mathrm{E}(Y) = \mu = b'(\theta)$ e $\mathrm{Var}(Y) = \phi b''(\theta)$ e, ϕ é um parâmetro de **dispersão** do modelo e seu inverso ϕ^{-1} é uma medida de **precisão**.

Os modelos lineares generalizados são definidos por (1) e pela componente sistemática

$$h(\mu_i) = \eta_i = \mathbf{x}_i^T \beta \tag{2}$$

em que η_i é o preditor linear, $\beta = (\beta_1, \ldots, \beta_p)^T$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ representa os valores de p covariáveis e $h(\cdot)$ uma função monótona e diferenciável, conhecida como função de ligação.

2.6 Modelos aditivos generalizados

O modelo aditivo generalizado (MAG) é uma extensão do modelo linear generalizado (MLG), em que o termo do modelo relacionado às covariáveis, $\sum_i x_{ij}\beta_j$, é substituído por $\sum_j g_j(X_j)$, com $g_j(X_j)$ denotando uma função semiparamétrica (i.e., cuja forma não é especificada) estimada através de funções de suavização (HASTIE; TIBSHIRANI, 1990). Com essa substituição, não é necessário assumir uma relação linear entre a variável resposta e as variáveis explicativas, como no MLG. 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, $\widehat{g}_j(X_j)$, também chamada de curva suavizada (smoother) ou função de suavização, em muitas situações, nada mais é do que algum tipo de média dos valores Y_i na vizinhança de um dado valor x_i mas, também, pode assumir diversas formas conforme o método de estimação utilizado.

A função de suavização permite então descrever a forma, e mesmo revelar uma possível ausência de linearidade nas relações estudadas, uma vez que não apresenta a estrutura rígida de uma função paramétrica. Os procedimentos de estimação para os MAGs são semelhantes àqueles adotados na estimação sob os MLGs, combinando métodos de suavização

com o método escore de Fisher[4].

Hastie e Tibshirani (1990) apresentam uma série de funções suavizadas, algumas mais refinadas do que a média móvel, a mais comumente utilizada e o caso mais simples de uma função suavizada, como os cubic smoothing splines ou o locally weighted running line smoother (loess) etc.

As duas considerações mais importantes para a escolha da função suavizada envolvem o tipo de função dos valores da variável resposta a ser estimada e um parâmetro de suavização h, o qual depende da função escolhida para a suavização. Este parâmetro, denominado banda, influencia o grau de suavidade na superfície estimada - quanto maior o valor de h, mais suavizada será a superfície estimada, produzindo estimativas com baixa variância, porém com viés alto.

Os modelos MAG tem a vantagem de permitir a estimação espacial do risco, por exemplo, fazendo-se o controle por fatores individuais de risco, os quais podem ser escritos da seguinte forma:

$$h(\mu) = \beta \boldsymbol{x} + g(s) \tag{3}$$

onde $\boldsymbol{x}$ é o vetor de covariáveis, β são os seus efeitos (parâmetros) e, g é uma função suavizada[5], porém desconhecida, das coordenadas espaciais s. Uma situação particular ocorreria se o risco for assumido constante na região de estudo, ou seja, $g(s) = 0$, o que faria com que o modelo (3) se reduzisse a um modelo de regressão usual.

2.7 Estudos caso-controle espaciais

Considere-se um tipo de estudo onde se tem dois tipos de eventos, por exemplo, a ocorrência ou não de algum tipo de doença em uma plantação, observada por determinado período. O total de plantas acometidas pela doença é uma variável do tipo binomial e, depende de diversas covariáveis, inclusive sua localização no espaço. Então, pode-se modelar o processo utilizando o método clássico de regressão logística, próprio para este tipo de distribuição. O que particulariza o contexto espacial é a forma de se incluir a localização dos pontos no modelo.

[4]coincidente com o método de Newton-Raphson no caso de funções de ligação canônica.

[5]única suposição feita.

Para a definição da medida de risco assume-se um desenho do tipo caso-controle espacial. No contexto deste trabalho, os casos serão os trabalhadores da região urbana de Piracicaba que se encontram em situação de precarização do trabalho e que sofreram algum tipo de acidente decorrente de seu(s) trabalho(s). Os controles serão os trabalhadores da mesma região, que sofreram algum tipo de agravo a saúde que não fora decorrente de acidente do trabalho.

Assumindo que as localizações de D (casos) e T (controles) sejam realizações independentes de dois processos de Poisson, com intensidades $\lambda_D(d)$ e $\lambda_T(t)$, respectivamente. Pode-se definir a função logarítmica do risco na localidade s a ser estimada em A, região de estudo, por:

$$\rho(s) = log\frac{\lambda_D(d)}{\lambda_T(t)} \tag{4}$$

e o objetivo na análise é investigar a variação espacial de $\rho(s)$ no espaço A.

A estimação da medida de risco tem aspectos muito semelhantes à estimação individual do risco. No entanto, ao invés de se avaliar somente as variações nas frequências de fatores individuais de risco entre casos e controles, avalia-se simultaneamente a variação na distribuição espacial de casos quando comparados à distribuição de controles (Shimakura et. al., 2001).

Supondo que casos e controles sejam amostras aleatórias de D e T com proporções desconhecidas q_D e q_T, respectivamente, em relação ao total de fato existente e, seja x_i, $i = 1, \ldots, n_D$, os n_D pontos observados de casos e, x_i, $i = n_{D+1}, \ldots, n_{D+T}$, os n_T pontos observados de controles, e seja Y_i uma variável associada ao ponto x_i, tal que $y_i = 1$ se $x_i \in D$ e $y_i = 0$, caso contrário.

Condicionada aos pontos x_i, considera-se que y_i são variáveis aleatórias independentes $Y_i \sim Bernoulli(p(s))$, onde $p(s) = P(Y_i = 1|X_i = s)$. Então,

$$p(s) = \frac{q_D\lambda_D(s)}{q_D\lambda_D(s) + q_T\lambda_T(s)} \tag{5}$$

Tem-se, então, que:

$$\log\left(\frac{p(s)}{1-p(s)}\right) = \rho(s) + c \tag{6}$$

onde, $c = \log \frac{q_D}{q_T}$, ou seja, c é simplesmente uma constante aditiva e, portanto, não modifica as características da distribuição espacial do risco sobre a região estudada (SHIMAKURA et al., 2001). E como extensão dessa abordagem, pode-se incorporar covariáveis não espaciais à eq. (6) por um modelo aditivo generalizado (MAG).

Kelsall e Diggle (1998) sugeriram que a inclusão de efeitos de covariáveis no modelo é dada por:

$$\log\left(\frac{p(s,x)}{1-p(s,x)}\right) = \beta x + g(s) \tag{7}$$

onde x é o vetor de covariáveis, β os seus efeitos e $g(s)$ uma função de suavização (por suposição), porém desconhecida, das coordenadas espaciais s. Novamente, uma situação particular pode ocorrer quando o risco for assumido constante na região de estudo, ou seja $g(s) = 0$, o modelo (7) reduziria-se a um modelo de regressão logística usual (HOSMER; LEMESHOW, 1989). Logo, o modelo (7) nada mais é do que um modelo de regressão logística estendido por uma componente aditiva $g(s)$ que, por suposição, tem variação suave no espaço estudado (SHIMAKURA et al., 2001).

O procedimento para estimar os efeitos β e a função $g(s)$ é baseado em métodos iterativos usuais de modelos aditivos generalizados (HASTIE; TIBSHIRANI, 1990), o qual pode ser resumido no seguinte algoritmo:

1. Para se obter as estimativas iniciais dos efeitos β faz-se $g(s) = 0$ e ajusta-se um modelo de regressão logística

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta x,$$

   (HOSMER; LEMESHOW, 1989).

2. Calcula-se

$$\hat{\eta}_i = \hat{\beta} x_i + \hat{g}(s),$$

$$\hat{p}_i = \frac{\exp(\hat{\eta}_i)}{1-\exp(\hat{\eta}_i)}$$

e

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}.$$

3. Ajusta-se um modelo aditivo da forma $Z = \beta x + g(s) + \varepsilon$ e:

   (a) Estima-se $g(s)$ usando regressão kernel ponderada (WAND; JONES, 1995) $u_i = z_i - \hat{\beta}x_i$ em s_i com pesos $w_i = \hat{p}_i(1-\hat{p}_i)$:

   $$\hat{g}(s) = \frac{\sum_{i=1}^{n} w_i K\left(\frac{s-s_i}{h}\right) u_i}{\sum_{i=1}^{n} w_i K\left(\frac{s-s_i}{h}\right)},$$

   onde $K(\cdot)$ é a função Kernel. Pode-se, na realidade, estimar $g(\cdot)$ de outras maneiras dentro do modelo MAG, por exemplo, por spline bidimensional. O parâmetro h conhecido como banda, influencia o grau de suavidade na superfície de risco estimada - quanto maior o valor de h, maior o alisamento.

   (b) Estima-se β por métodos de mínimos quadrados ponderados de $z_i - \hat{g}(s_i)$ em u_i com pesos w_i.

   (c) Repete-se os passos $(a.)$ e $(b.)$ até a convergência das estimativas.

4. Repete-se os passos (2.) e (3.) até a convergência das estimativas.

2.8 Modelos estruturados aditivamente

Nas situações onde a classe de modelos MLG deixam de ser adequados, utiliza-se uma classe de modelos de regressão estruturados aditivamente. Sua utilização é devida à flexibilidade existente nela (FAHRMEIR; TUTZ, 2001).

Seja uma variável aleatória Y_i com distribuição de probalidade possível de ser escrita na forma da família exponencial, onde a média μ_i é ligada ao preditor estruturado aditivamente η_i através de uma função de ligação $g(\cdot)$, tal que $g(\mu_i) = \eta_i$. O preditor pode acomodar diferentes efeitos de várias covariáveis da forma aditiva

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{8}$$

Aqui, os $f^{(j)}(\cdot)$ são funções desconhecidas das covariáveis u, os β_k representam efeitos lineares das covariáveis z e os ε_i são termos não estruturados. Este modelo pode ser aplicado de diversas formas, dependendo da forma que as funções $f^{(j)}$ podem assumir. Os modelos Gaussianos latentes são um subconjunto de todos os modelos Bayesianos estruturados aditivamentes, onde se supõe uma priori Gaussiana para α, $f^{(j)}$, β_k e ε_i.

Rue, Martino e Chopin (2009) propuseram uma abordagem para fazer inferência Bayesiana aproximada em modelos Gaussianos latentes, a qual é denominada de INLA (INTEGRATED NESTED LAPLACE APPROXIMATIONS), os quais mostram que esta abordagem é extremamente rápida em termos de implementação, principalmente, por recorrer à utilização de algoritmos para matrizes esparsas, inerentes a essa classe de modelos. Os autores ainda mostram que, esta abordagem é superior à MCMC (MARKOV CHAIN MONTE CARLO) em termos de acurácia e tempo computacional. Eles, ainda descrevem como utilizar aproximações para derivar ferramentas de teste para o erro de aproximação, aproximar posterioris marginais, cálculos de medidas de diagnóstico como, critério de informação da Deviance (DIC) e outras medidas preditivas Bayesianas.

Nas aplicações de modelos estruturados aditivamente, o modelo final consistirá na soma de vários componentes, tais como, efeito espacial, efeitos aleatórios, efeitos lineares e não lineares de covariáveis.

A notação utilizada aqui será a mesma de Bonat (2010), cujo autor discute métodos de inferência Bayesiana aproximada para modelos Gaussianos latentes em dados espaço-temporais.

Considerando apenas modelos com efeitos principais, seja Y_i a variável observada na área i ($i = 1, \ldots, n$). Assumindo que a distribuição de probabilidade de Y_i tem distribuição pertencente à família exponencial, com média μ_i e, possivelmente, parâmetro de escala ou dispersão ϕ, que pode ou não depender do parâmetro de média. Considerando-se a função de ligação $g(\cdot)$, tem-se que $g(\mu_i) = \eta_i$ é o preditor linear com a seguinte decomposição

$$\eta_i = \alpha + \phi_i + \varphi_i \tag{9}$$

onde α é o nível médio do processo e ϕ_i e φ_i, representam desvios da média geral para a área i, a qual possui e não possui estrutura espacial, respectivamente.

Define-se para (9), segundo blocos $\varphi = (\varphi_1, \ldots, \varphi_n)^T$ e $\phi = (\phi_1, \ldots, \phi_n)^T$, distribuições a priori Gaussianas com média zero e matriz de precisão $k\mathbf{K}$, onde k é um escalar desconhecido e, $\mathbf{K}$ é uma matriz de estrutura conhecida, a qual será diferente para cada bloco, no intuito de descrever, a priori, suposições diferentes para o relacionamento entre os parâmetros de cada bloco.

Para o bloco estruturado espacialmente $\boldsymbol{\varphi}$, a matriz de estrutura é uma simples autoregressão Gaussiana (BESAG; YORK; MOLLIÉ, 1991). Seus elementos de estrutura $\mathbf{K}_\varphi$ são $k_{ij} = -1$ para áreas vizinhas ($i \sim j$) e, elementos k_{ii} igual ao número de áreas geograficamente contíguas a área i. Os demais elementos de $\mathbf{K}_\varphi$ são zero. A distribuição a priori para o componente $\boldsymbol{\varphi}$ pode ser escrita como

$$\pi(\boldsymbol{\varphi}|k_\varphi) \propto \exp\left(-\frac{k_\varphi}{2}\sum_{i\sim j}(\varphi_i - \varphi_j)^2\right) \tag{10}$$

Para a heterogeneidade espacial não estruturada, toma-se como matriz de estrutura a matriz identidade $\mathbf{K}_\phi = I$.

Seja $\pi(\cdot|\cdot)$ a densidade condicional de seus argumentos e seja $\boldsymbol{x}$ a representação das n variáveis Gaussianas η_i, α, $f^{(j)}$ e β_k. A densidade $\pi(\boldsymbol{x}|\theta_1)$ é Gaussiana com média zero (por suposição), e matriz de precisão $\boldsymbol{Q}(\theta_1)$. Denote, também, por $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ a densidade Gaussiana $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ainda na configuração de $\boldsymbol{x}$.

A distribuição para as n variáveis observadas $\boldsymbol{y} = y_i : i \in \mathbf{I}$ é denotada por $\pi(\boldsymbol{y}|\boldsymbol{x}, \theta_2)$ e supõe-se que y_i são condicionalmente independentes $\boldsymbol{x}$ e θ_2.

Quando é suposta independência condicional de $\pi(\boldsymbol{y}|\boldsymbol{x}, \theta_2)$, sua distribuição se resume ao produto da distribuição suposta para a variável resposta, ou seja, a função de verossimilhança.

Como forma de simplificação, denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$ com $dim(\boldsymbol{\theta}) = m$. A distribuição a posteriori para uma matriz $\boldsymbol{Q}(\boldsymbol{\theta})$ não singular é dada por

$$\begin{aligned}\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i\in I}\pi(y_i|x_i,\boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})||\boldsymbol{Q}(\boldsymbol{\theta})||^{n/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}+\sum_{i\in I}\log\pi(y_i|x_i,\boldsymbol{\theta})\right) \qquad (11)\end{aligned}$$

A imposição de restrições lineares, quando necessário, são denotadas por $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{e}$ para uma matriz $\boldsymbol{A}$ de dimensão $k\times n$ de posto k. O objetivo principal é aproximar as marginais a posteriori $\pi(x_i|\boldsymbol{y})$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ e $\pi(\boldsymbol{\theta}_j|y)$. Muitos dos modelos Gaussianos latentes satisfazem duas propriedades básicas, as quais serão assumidas neste trabalho. A primeira, é que o campo latente $\boldsymbol{x}$, o qual, em geral, tem dimensão grande, admite propriedades de independência condicional. Assim, o campo latente é um Campo Aleatório Markoviano Gaussiano (CAMG) com matriz de precisão $\boldsymbol{Q}(\boldsymbol{\theta})$ (RUE; HELD, 2005). Isto significa que se pode usar métodos numéricos para matrizes esparsas, os quais são muito mais rápidos que os outros métodos (RUE; HELD, 2005). A segunda propriedade é que o número de hiperparâmetros m é pequeno, ou seja, $m\leq 6$. Ambas propriedades são muito úteis para produzir inferência rápida, mesmo que existam exceções (EIDSVIK; MARTINO; RUE, 2009).

A abordagem INLA trabalha usando o fato de que a distribuição marginal a posteriori de interesse pode ser escrita como

$$\pi(x_i|\boldsymbol{y})=\int\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\partial\boldsymbol{\theta} \quad \text{e} \quad \pi(\boldsymbol{\theta}_j|\boldsymbol{y})=\int\pi(\boldsymbol{\theta}|\boldsymbol{y})\partial\boldsymbol{\theta}_{-j} \qquad (12)$$

O ponto chave desta tipo de abordagem é a utilização desta forma para construir aproximações aninhadas, do tipo

$$\tilde{\pi}(x_i|\boldsymbol{y})=\int\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\partial\boldsymbol{\theta} \quad \text{e} \quad \tilde{\pi}(\boldsymbol{\theta}_j|\boldsymbol{y})=\int\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\partial\boldsymbol{\theta}_{-j} \qquad (13)$$

Como notação, $\tilde{\pi}(\cdot|\cdot)$ é a densidade condicional aproximada de seus argumentos. Aproximações para $\pi(x_i|\boldsymbol{y})$ são calculadas aproximando $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ e $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$, e usando integração numérica (soma finita) para integrar fora $\boldsymbol{\theta}$. A integração é possível quando a dimensão de $\boldsymbol{\theta}$ é pequena, em geral menor ou igual a 6.

A abordagem INLA tem base na seguinte aproximação $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ para a distribuição marginal a posteriori de $\boldsymbol{\theta}$.

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{14}$$

onde, $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ é a aproximação Gaussiana para a condicional completa de $\boldsymbol{x}$ o que caracteriza a aproximação como a de Laplace e, $\boldsymbol{x}^*(\boldsymbol{\theta})$ é a moda da distribuição condicional completa de $\boldsymbol{x}$ para um dado $\boldsymbol{\theta}$. A equação (14) é válida em um ponto apenas e, portanto, para se obter a aproximação para a distribuição completa, é necessário avaliá-la para um conjunto de valores de $\boldsymbol{\theta}$. A proporcionalidade em (14) segue do fato que a constante normalizadora de $\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ é desconhecida. A distribuição $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ tende a ser muito diferente da distribuição Gaussiana, o que sugere que uma aproximação Gaussiana direta para $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ não é muito acurada. Rue e Martino (2007) utilizaram (14) para aproximar distribuições marginais a posteriori para muitos modelos Gaussianos latentes, os quais concluíram que $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ é acurada, mesmo na situação de execução de um longo MCMC. Para as distribuições marginais a posteriori do campo latente, é proposto começar por $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$, ou seja,

$$\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y}) \quad \text{com distribuição} \quad N(x_i;\mu_i(\boldsymbol{\theta}),\sigma_i^2(\boldsymbol{\theta})). \tag{15}$$

Para esta notação, $\mu(\boldsymbol{\theta})$ representa o vetor de médias para a aproximação Gaussiana e, $\sigma^2(\boldsymbol{\theta})$ o vetor de variâncias marginais, cuja aproximação pode ser integrada numericamente com relação a $\boldsymbol{\theta}$ (ver Eq. 13) para obter aproximações para as marginais de interesse do campo latente,

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y}) \times \Delta_k. \tag{16}$$

A soma é realizada sobre os valores de $\boldsymbol{\theta}$ com pesos Δ_k. Rue e Martino (2007) mostraram que a distribuição marginal a posteriori para $\boldsymbol{\theta}$ foi acurada, enquanto o que o erro na aproximação Gaussiana (15) foi grande. Em particular, (15) pode apresentar erro na locação e/ou assimetria (skewness). As dificuldades de Rue e Martino (2007) foram em detectar os x_i nas quais a aproximação era menos acurada e a falta de habilidade para melhorar a aproximação nestas localizações. Além disso, não conseguiam controlar o erro de aproximação e escolher os pontos de integração θ_k de uma forma adaptativa e automática.

Rue, Martino e Chopin (2009) resolveram os problemas encontrados em Rue e Martino (2007) e, apresentaram abordagem completa para a inferência aproximada em modelos Gaussianos latentes, a qual fora nomeada de Integrated Nested Laplace Approximations

(INLA). Como principal mudança, teve-se novamente a aplicação da aproximação de Laplace, agora em $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$. Os autores apresentaram, ainda, uma alternativa rápida, a qual corrige a aproximação Gaussiana (15) para o erro de locação e a falta de simetria com custo extra moderado. Tais correções são obtidas por expansão em séries de Laplace. Esta alternativa rápida é a primeira escolha natural por causa do seu baixo custo computacional e alta acurácia. Os autores demonstram como várias aproximações podem ser usadas para se obter ferramentas de teste da aproximação, aproximar distribuições marginais a posteriori para um subconjunto de $\boldsymbol{x}$ e, calcular diversas medidas de interesse, tais como verossimilhança marginal, Critério de Informação da Deviance (DIC) e várias medidas preditivas Bayesianas.

A abordagem INLA para aproximar distribuições marginais a posteriori para campos latentes Gaussianos $\pi(x_i|\boldsymbol{y})$, $i = 1,\ldots,n$ pode ser feita em três passos. O primeiro passo aproxima a distribuição marginal a posteriori de $\boldsymbol{\theta}$ através da aproximação de Laplace (14). O segundo passo calcula a aproximação de Laplace para $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$, para valores selecionados de $\boldsymbol{\theta}$, com o intuito de melhorar a aproximação Gaussiana (15). O terceiro passo combina os dois anteriores usando a integração numérica (16).

O primeiro passo da abordagem INLA é o de calcular uma aproximação para a distribuição a posteriori marginal de $\boldsymbol{\theta}$ (ver Eq. 14). Seu denominador é a aproximação Gaussiana para a distribuição condicional completa de $\boldsymbol{x}$ (BONAT, 2010). O principal uso de $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ é integrar fora a incerteza com relação a $\boldsymbol{\theta}$ quando se aproxima a distribuição marginal a posteriori de x_i (ver Eq. 16). Em Bonat (2010) são descritos em detalhes como se explorar a distribuição marginal a posteriori para encontrar bons pontos para o cálculo da integração numérica.

A aproximação da distribuição marginal a posteriori para θ_j pode ser realizada diretamente de $\tilde{\pi}(\boldsymbol{\theta})$ através de integração numérica. O custo computacional para isto é relativamente alto porque é necessário calcular $\tilde{\pi}(\boldsymbol{\theta})$ para um grande número de configurações. Uma abordagem menos custosa é utilizada através do uso dos pontos obtidos nos passos 1-3 para construir um interpolante para $\log\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, e calcular marginais através de integração numérica vinda dele.

Como passo seguinte, é preciso providenciar aproximações acuradas para a distribuição a posteriori marginal dos $x_i's$ condicionada nos valores selecionados de $\boldsymbol{\theta}$, os quais são

pontos ponderados $\boldsymbol{\theta}_k$ a serem utilizados em (16). Será discutido de forma sucinta a aproximação de Laplace para $\tilde{\pi}(x_i|\boldsymbol{y}, \theta_k)$. Em Bonat (2010) há uma discussão mais ampla e, uma aproximação Gaussiana para a distribuição a posteriori.

Bonat (2010) afirma que a forma natural de melhorar a aproximação Gaussiana, discutida em detalhes no mesmo, é calcular a aproximação de Laplace

$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(x, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=x^*_{-i}(x_i, \boldsymbol{\theta})} \tag{17}$$

O denominador $\tilde{\pi}_{GG}$ é a aproximação Gaussiana para $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$ e x^*_{-i} corresponde à configuração modal. A densidade $\tilde{\pi}_{GG}$ precisa ser recalculada em (17) para cada valor de x_i e $\boldsymbol{\theta}$, pois sua matriz de precisão depende de x_i e $\boldsymbol{\theta}$. Tal procedimento é custoso porque requer n fatorações da matriz de precisão completa. Rue, Martino e Chopin (2009) propuseram duas modificaçoes em (17) no intuito de superar esta dificuldade computacional, as quais são discutidas de forma detalhada em Bonat (2010).

O cálculo da verossimilhança marginal é útil na comparação de modelos, assim como, o fator de Bayes é definido pela razão de verossimilhanças marginais entre dois modelos competidores. Ela é expressa por

$$\tilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x} = \boldsymbol{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{18}$$

em que, $\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.

Para avaliar os modelos ajustados utiliza-se como principal medida o critério de informação da Deviance (SPIEGELHALTER et al., 2001), o qual é usual entre os modelos hierárquicos e, em geral, bem definido para prioris impróprias. Tal critério tem como principal aplicação a seleção de modelos bayesianos, no entanto, também é aplicado na obtenção do número efetivo de parâmetros do modelo ao qual é aplicado. Sua forma analítica é

$$D(\boldsymbol{x}, \boldsymbol{\theta}) = -2 \sum_{i \in I} \log \pi(y_i|x_i, \boldsymbol{\theta}) + \text{constante}. \tag{19}$$

cuja medida é definida como duas vezes a média da *deviance* menos a deviance para a média. O número efetivo de parâmetros é a média da deviance menos a deviance da média.

3 MATERIAL E MÉTODOS

3.1 Material

Para a realização deste trabalho, utilizou-se dos dados sobre acidentes de trabalhos, os quais são oriundos do Projeto temático FAPESP 2006/05920-7: Estimabilidade de Medidas de Associação e de Risco em Estudos Caso-Controle Espaciais, coordenado pelo Professor Dr. Ricardo Cordeiro, da Faculdade de Ciências Médicas da UNICAMP. Estudo este que tem como unidade de análise os trabalhadores em situação de precarização de trabalho na região urbana do município de Piracicaba/SP cujas ocorrências de acidente de trabalho estão georreferenciadas em seus respectivos locais ocorridos, para os casos e, os locais de trabalho para os controles.

A população fonte de casos é composta pelos trabalhadores precarizados que trabalham na área urbana de Piracicaba e que sofreram acidente de trabalho no período entre setembro de 2006 e agosto de 2007.

A população fonte de controle foi alocada a partir da população fonte de casos, ou seja, é composta pelos trabalhadores em situação de precarização do trabalho, que trabalhavam na área urbana de Piracicaba no período estudado mas, que não sofreram acidente por motivo de trabalho e, sim por outro motivo qualquer ou que acompanhavam os indivíduos considerados casos no momento de atendimento destes no pronto-socorro. Para estes, foram localizados em mapa digital da área estudada os respectivos locais de trabalhos, enquanto para os casos, os locais da ocorrência do acidente de trabalho. Utilizou-se uma amostra como controle para se tentar refletir, o máximo possível, como a população da área estudada se distribuía no espaço.

Os dados foram coletados em alguns pronto-socorros de Piracicaba, proporcionalmente à quantidade histórica de atendimentos a acidentes de trabalho de cada um, conforme dados do SIVAT - Sistema de Vigilância de Acidentes do Trabalho de Piracicaba de 2001 a 2005.

Foram entrevistados 2.451 trabalhadores que atendessem as condições do estudo: ser trabalhador precarizado (sem carteira assinada ou terceirizado ou doméstico ou que trabalha na rua), maior de 16 anos, que trabalhava na região urbana de Piracicaba, dos quais 819 foram atendidos em virtude de acidente de trabalho (casos) e 1.632 que procuraram o serviço por

causas outras diferentes de acidente de trabalho (controle).

Devido a alguns problemas encontrados pela equipe que coletou os dados, no processo de revisão final, e validação dos dados coletados, considerou-se 810 casos e 1.620 controles, o que totaliza 2.430 trabalhadores entrevistados, porém, a análise aqui realizada considera todos os dados.

O georreferenciamento no mapa da área urbana de Piracicaba referentes ao endereço de cada ocorrência de acidente, assim como, o georreferenciamento do endereço dos controles (não-casos) está ilustrado na figura (1) abaixo.

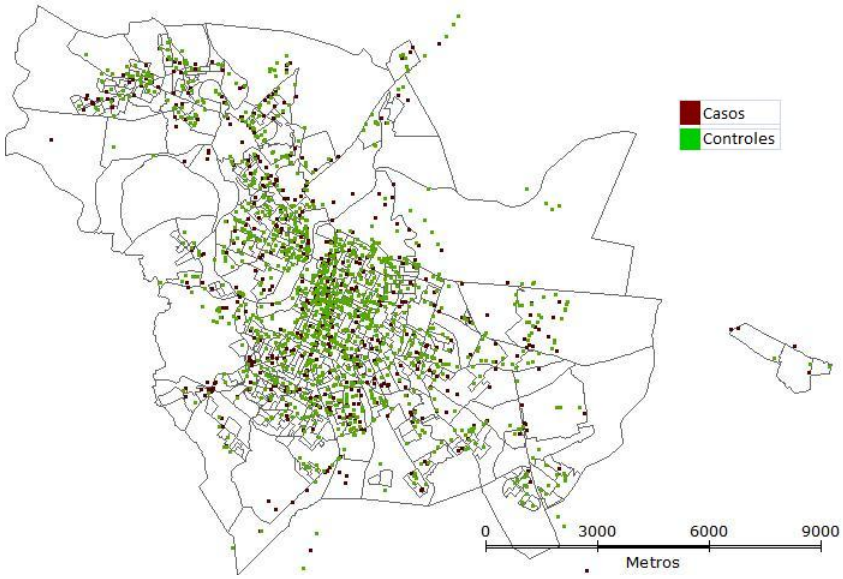

Figura 1 - Mapa de casos e controles de acidente de trabalho, segundo setores censitários

3.2 Métodos

O risco de ocorrer algum "evento" pode ser maior em certos locais ou regiões do que em outros. A incidência de um determinado agravo à saúde também pode ter riscos diferentes entre regiões. Tais eventos também podem ser função do espaço.

Nos anos 90 começou-se a desenvolver uma especialização dos estudos caso-controle que incorporam explicitamente o espaço ao conjunto de covariáveis que modelam o risco. São os chamados estudos caso-controle espaciais, que então passam a contribuir com a incorporação de métodos de análise espacial de dados.

Na epidemiologia espacial, um problema comumente ocorrido é o de determinar se os casos de uma certa doença têm algum tipo de associação espacial. Isto pode ser verificado comparando a distribuição espacial da localização dos casos a um conjunto de controles tomados aleatoriamente da população estudada.

Devido ao grande desenvolvimento de estudos caso-controle, bem como a contribuição analítica que os sistemas de informação geográfica e os métodos de análise espacial de dados trazem para a epidemiologia, o projeto no qual este trabalho teve origem tem como objetivos, em estudos caso-controle espaciais:

1. desenvolver modelos computacionais que simulem, para um conjunto de parâmetros demográficos, geográficos e epidemiológicos pré-definidos, cenários de dinâmica espacial de ocorrência de casos e o processo de execução de estudos caso-controle espacial na população.

2. desenvolver métodos para modelar e estimar os componentes do modelo para o risco.

A análise de um processo pontual está focada na distribuição espacial dos eventos observados e na realização de inferências acerca do processo que os gerou. No caso, há dois principais interesses: a distribuição dos eventos no espaço estudado e a existência de uma possível interação entre eles.

Como início de análise de padrões de pontos tem-se como alternativa a estimação da intensidade pontual do processo na região estudada. Para isto, ajusta-se uma função bidimensional sobre os eventos considerados, no caso, ocorrência de acidente de trabalho entre trabalhadores em situação de precarização do trabalho, compondo uma superfície cujo valor será proporcional à intensidade de amostras por unidade de área (DRUCK et al., 2004).

A função ajustada faz uma contagem dos pontos dentro de uma região de influência, ponderando-os pela distância de cada um à localização de interesse, como mostrado de forma esquemática na figura (2) abaixo.

O estimador de kernel é definido pela seguinte função (DRUCK et al., 2004):

$$\hat{\lambda}_r(u) = \frac{1}{r^2} \sum_{i=1}^{n} K\left(\frac{d(u_i, u)}{r}\right), d(u_i, u) \leq r \tag{20}$$

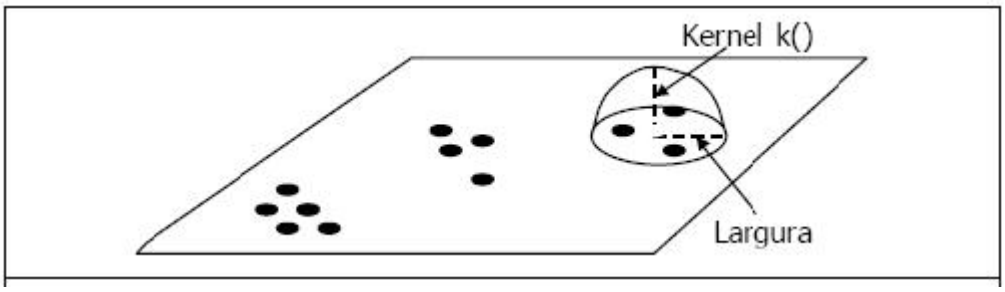

Figura 2 - Esquema do estimador de *kernel*

em que $r \geq 0$ define o raio de influência, o qual define a vizinhança do ponto a ser interpolado e controla a suavização da superfície gerada; $d(u_i, u)$ é a distância entre o i-ésimo ponto observado e o ponto a ser estimado e $k(\cdot)$ é uma função de interpolação (kernel), que em geral, é de terceira ou quarta ordem.

A ocorrência de algum evento de interesse pode ocorrer de forma agregada, simplesmente porque ocorrem com mais frequência em áreas com maior densidade populacional. Portanto, os métodos que não levam isso em conta podem levar a conclusões equivocadas quanto à distribuição espacial de tais eventos e, uma possibilidade de corrigir isso, seria levar em consideração a distribuição populacional na região de estudo, o que nem sempre se tem à disposição.

Uma maneira de contornar esse tipo de problema, se a densidade populacional for a única fonte de variação, é observar o padrão pontual de um grupo controle.

Uma alternativa inicial para verificação espacial do padrão pontual seria a utilização da função K, a qual é eficiente para detectar padrões de agrupamentos de eventos, assim como, eventos regularmente distribuídos. Esta função foi proposta por Ripley (1977) e é definida formalmente como:

$$K(t) = \frac{2\pi}{\lambda^2} \int_0^t \lambda^2(x) x dx. \tag{21}$$

Uma função K corrigida para o tipo de problema acima, conhecida como função K cruzada, pode ser utilizada, a qual é definida como

$$K(d, t) = K_D(d) K_T(t) \tag{22}$$

onde D denota os casos e T os controles. Este procedimento é válido desde que os casos e controles não sejam pareados espacialmente, ou seja, desde que eles ocorram de forma independente.

A função K Cruzada indicará se casos e controles se distribuem de forma agregada, aleatória ou regular na área analisada levando em consideração a distribuição populacional.

Para uma análise inferencial, através da aplicação de um modelo tem-se que "ao se estabelecer um modelo de regressão buscando relacionar uma variável resposta a variáveis independentes, um dos pressupostos básicos da estatística, da independência entre amostras, e pouco realístico: na verdade, nos dados espaciais a dependência está presente em todas as direções e fica mais fraca à medida em que aumenta a dispersão na localização dos dados."(CARVALHO; SOUZA-SANTOS, 2005).

Os MAG, assim como os MLG, generalizam a relação do valor esperado de uma variável resposta com covariáveis que são submetidas a funções de suavização $g_j(\cdot)$. Temos então,

$$\eta(E(Y|X_1,\ldots,X_p)) = \beta_0 + \sum_{j=1}^{p} g_j(X_j) \tag{23}$$

A função de suavização $g_j(\cdot)$ contempla as covariáveis do modelo, podendo incluir as coordenadas geográficas de pontos no espaço. Desta forma pode-se incorporar um componente espacial ao modelo fazendo com que ele se torne mais informativo quando o objetivo é descrever padrões espaciais.

Dos dados coletados dos indivíduos (casos e controles) foram obteve-se as seguintes informações: sexo, idade, escolaridade, ocupação (grupo CBO[6]), ramo de atividade (grupo CNAE[7]) e risco referido pelos indivíduos de sofrer acidente de trabalho. Além dessas informações, para os indivíduos considerados casos fora obtido o tipo, a gravidade, o local e as características do acidente sofrido. E, para os controles fora obtido o(s) local(is) de trabalho.

Para a utilização da metodologia de modelos Gaussianos latentes, através da abordagem INLA, a qual será aplicada nos dados de área dos 507 setores censitários da área

[6]Classificação Brasileira de Ocupações.

[7]Classificação Nacional de Atividades Econômicas.

urbana de Piracicaba/SP, cuja notação utilizada aqui será a mesma do tópico (2.7), que fora replicada de Bonat (2010), cujo autor discute métodos de inferência Bayesiana aproximada para modelos Gaussianos latentes em dados espaço-temporais.

4 RESULTADOS E DISCUSSÃO

Nesta seção, com o tratamento dos dados sobre acidentes de trabalho, será mostrado os resultados obtidos e suas respectivas análises. O intuito é mostrar que a inclusão da componente espacial no processo de modelagem, há um ganho significativo na explicação do fenômeno estudado sob duas abordagens: uma por processo pontual e outra por dados de área.

4.1 Análise descritiva

Com o intuito de conhecer um pouco o conjunto de dados, realizou-se uma análise exploratória separada por casos e controle segundo as variáveis a serem trabalhadas. Abaixo, a Tabela (1) apresenta a distribuição por sexo entre casos e controles, considerando o total de 2.451 observações[8].

Tabela 1 - Distribuição do sexo segundo casos e controles - Piracicaba - 2006/2007

Sexo	Casos		Controles		Total	
	Absoluto	Percentual	Absoluto	Percentual	Absoluto	Percentual
Masculino	670	81,8	1.049	64,3	1.719	70,1
Feminino	149	18,2	583	35,7	732	29,9
Total	819	100,0	1.632	100,0	2.451	100,00

Observa-se que o número de pessoas do sexo masculino é sempre maior. Em termos percentuais, essa diferença é muito maior entre os casos. A Tabela (2) abaixo mostra a distribuição etária dos indivíduos participantes da pesquisa segundo a divisão entre casos e controles.

Observa-se que os casos tem maior participação na faixa de 20 a 29 anos, enquanto os controles na faixa de 30-39. De maneira geral, os indivíduos encontram-se em sua maioria nas faixas de 20 a 39 anos.

[8]Lembrando que no processo de modelagem foram considerados apenas 2.430 observações, devido a problemas encontrados durante a consistência dos dados.

Tabela 2 - Distribuição etária segundo casos e controles - Piracicaba - 2006/2007

Faixa etária	Casos		Controles		Total	
(anos)	Absoluto	Percentual	Absoluto	Percentual	Absoluto	Percentual
10-19	72	8,8	68	4,2	140	5,71
20-29	292	35,7	445	27,3	737	30,06
30-39	220	26,9	514	31,5	734	29,94
40-49	141	17,2	382	23,4	523	21,33
50-59	69	8,4	173	10,4	242	9,87
60-69	21	2,6	42	2,6	63	2,6
70-79	4	0,5	8	0,5	12	0,5
Total	819	100,0	1.632	100,0	2.451	100,00

A Tabela (3) a seguir mostra a distribuição da escolaridade, em anos, dos indivíduos segundo casos e controles.

Um pouco mais de um quarto dos indivíduos possui mais de 11 anos de estudo, tanto entre os casos como entre os controles. Observa-se também, um maior percentual de pessoas sem estudo entre os casos do que entre os controles.

A tabela (4) a seguir mostra a distribuição do risco de sofrer acidente de trabalho referido pelos trabalhadores entrevistados, tanto casos quanto controles.

Existem suspeitas de que ao referir um risco o trabalhador tende a elevá-lo, o que causa uma possível superestimação do escore atribuído. Isto é possível observar na tabela (4) que escores do risco acima de cinco tem maiores frequências.

Para completar a primeira parte de análise descritiva, uma tabela com a distribuição da classificação quanto à gravidade do acidente sofrido entre os trabalhadores acidentados (casos), aferido pelo médico que os atendeu nos Pronto-socorros respectivos, de acordo com uma escala pré-definida na ficha de atendimento do acidentado(gravidade crescente entre 1 e 4) é mostrada na Tabela (5) a seguir.

Dos dados da Tabela (5) percebe-se que mais da metade dos casos atendidos por acidente de trabalho foram classificados como de menor gravidade (gravidade 1), representando cerca de 71%.

Tabela 3 - Distribuição da escolaridade segundo casos e controles - Piracicaba - 2006/2007

Escolaridade	Casos		Controles		Total	
(em anos)	Absoluto	Percentual	Absoluto	Percentual	Absoluto	Percentual
0	27	3,3	25	1,5	52	2,1
1	15	1,8	38	2,3	53	2,1
2	29	3,5	35	2,1	64	2,6
3	23	2,8	57	3,5	80	3,2
4	83	10,1	214	13,1	297	12,2
5	88	10,7	150	9,2	238	9,7
6	50	6,1	103	6,3	153	6,2
7	46	5,6	83	5,1	129	5,3
8	143	17,5	267	16,4	410	16,8
9	26	3,2	38	2,3	64	2,6
10	43	5,3	54	3,3	97	3,9
11	211	25,8	442	27,1	653	26,6
Sup. Incomp.	27	3,3	63	3,9	90	3,7
Sup. Comp.	8	1,0	63	3,9	71	2,9
Total	819	100,0	1.632	100,0	2.451	100,00

Ainda como parte da análise exploratória, uma suavização da região pelo método kernel na área de estudo foi realizada com o intuito de se ter uma idéia da superfície de intensidade.

O método de suavização por kernel é uma alternativa simples para analisar o comportamento de padrões de pontos e muito útil para fornecer uma visão geral da distribuição de primeira ordem das ocorrências de acidente de trabalho (evento de interesse). Ele permite estimar a intensidade pontual do processo em toda a área estudada. Ele consiste em se contar a quantidade de pontos em cada região, utilizando uma janela móvel e com formatos que permitem que pontos mais próximos do ponto a ser estimado tenham maior influência na estimação da densidade (PICHARSKI et al., 2009).

Abaixo a Figura (3) mostra a suavização por kernel para os casos (acidentados

Tabela 4 - Distribuição do risco referido segundo casos e controles - Piracicaba - 2006/2007

Risco	Casos		Controles	
Referido	Absoluto	Percentual	Absoluto	Percentual
0	25	3,1	79	4,8
1	18	2,2	38	2,3
2	39	4,8	74	4,5
3	42	5,1	105	6,4
4	31	3,8	74	4,5
5	165	20,2	332	20,3
6	59	7,2	118	7,2
7	66	8,1	143	8,8
8	134	16,4	309	18,9
9	54	6,6	82	5,0
10	186	22,6	278	17,0
Total	819	100,0	1.632	100,0

Tabela 5 - Distribuição da classificação da gravidade do acidente - Piracicaba - 2006/2007

Gravidade	Frequência	Percentual
1	581	70,9
2	215	26,3
3	22	2,7
4	1	0,1
Total	819	100,0

em decorrência do trabalho), para os controles (acidentados não decorrentes do trabalho) e para o risco de sofrer acidente de trabalho.

Para verificar através de uma medida de segunda ordem como os dois processos pontuais aqui considerados (casos e controle) se distribuem conjuntamente, utiliza-se a função K cruzada, a qual é mostrada na Figura (4) abaixo. As linhas pontilhadas representam um envelope de simulações via Monte Carlo. A linha contínua representa os valores empíricos da

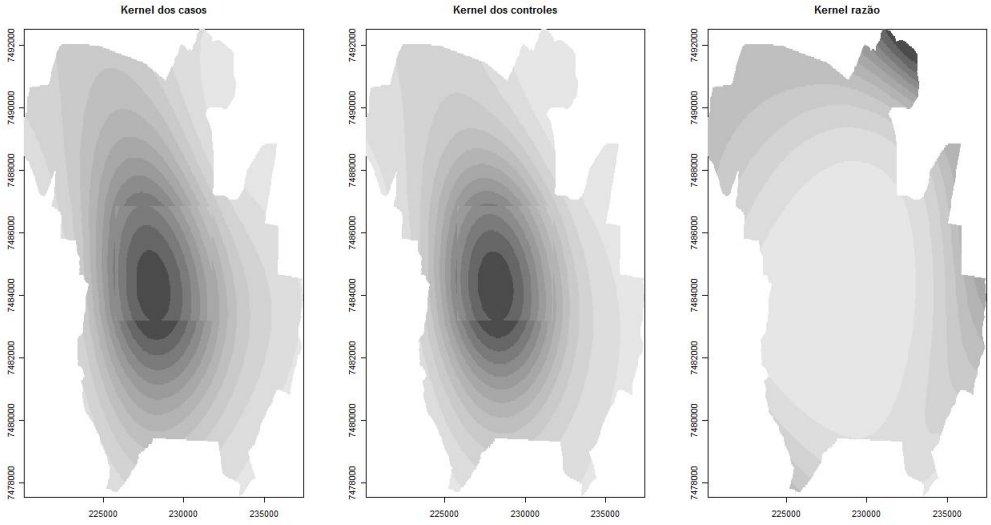

Figura 3 - Mapa de *kernel*

função.

Pode-se observar que a linha que representa os valores empíricos está completamente acima do envelope de simulação. Pode-se então concluir que os casos de acidentes de trabalho na área urbana de Piracicaba estão espacialmente agregados, o que já era esperado devido às características dos indivíduos que compõem a amostra da pesquisa.

Durante o tratamento dos dados encontrou-se coordenadas com valores zero totalizando um total de quarenta registros, restando 2.411 registros válidos para dar continuidade na análise.

Parte desses registros com problemas são aqueles que foram desconsiderados no processo de validação dos dados, realizado pela equipe que fez a coleta da amostra.

Observou-se também a ocorrência de mais de um caso e/ou mais de um controle por localização geográfica e para contornar possíveis problemas durante a análise inferencial utilizou-se do recurso de deslocar esses pontos coincidentes adicionando uma constante às coordenadas. A constante considerada foi um raio de cinquenta metros em torno dos pontos em que ocorreu mais de uma ocorrência. Nos pontos onde não ocorreu mais de uma ocorrência não somou-se a constante às coordenadas. Decidiu-se utilizar este recurso, pois dificilmente ao somar tal constante, as novas coordenadas seriam coincidentes.

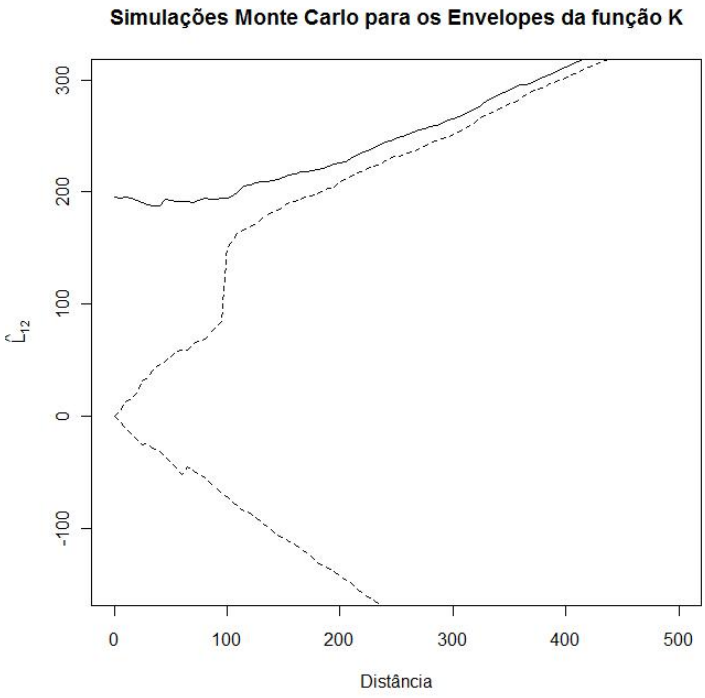

Figura 4 - Função K cruzada com envelopes de simulação via MCMC

Com o intuito de verificar se aconteceu alguma mudança nos dados, uma nova análise exploratória considerando as novas coordenadas foi realizada, e os resultados são mostrados a seguir.

De forma geral, não obteve-se resultados diferentes dos obtidos anteriormente para os mapas de kernel, o qual mostra que a intensidade é praticamente a mesma para o risco de sofrer acidente de trabalho nem para a função K cruzada, a qual identificou que os dados se distribuem de forma agregada, conforme encontrado anteriormente.

4.2 Análise inferencial

Para dar início à análise inferencial dos dados, a qual incluirá métodos de estimação, modelos etc., um modelo de regressão logística múltipla tendo como variável resposta o status do indivíduo (caso ou controle) e covariáveis, as variáveis citadas anteriormente na Seção 3.2 foi ajustado com o seguinte resultado mostrado na tabela (6), a qual apresenta o modelo MLG ajustado, inicialmente testado, para as variáveis listadas acima.

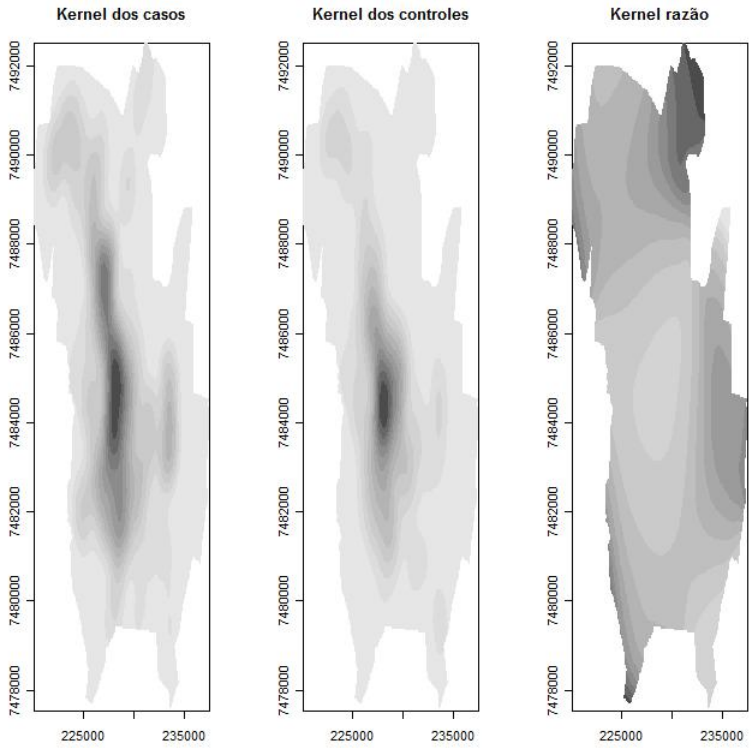

Figura 5 - Mapa de kernel após o deslocamento dos pontos coincidentes

Tabela 6 - Valores estimados para o MLG inicial

Coeficientes	Estimativa	Erro Padrão	Estatística de teste	Valor p
Intercepto	-0.79960	0.27124	-2.948	0.00320
idade	-0.02096	0.00418	-5.015	<0.0001
sexo	0.86624	0.12676	6.834	<0.0001
anos escolaridade	-0.02269	0.01323	-1.716	0.08625
risco referido	0.04301	0.01641	2.621	0.00878
ctps	0.64142	0.11227	5.713	<0.0001
na rua	-0.32507	0.10687	-3.042	0.00235
terceirizado	0.03453	0.12912	0.267	0.78915
doméstico	-0.32051	0.18220	-1.759	0.07856

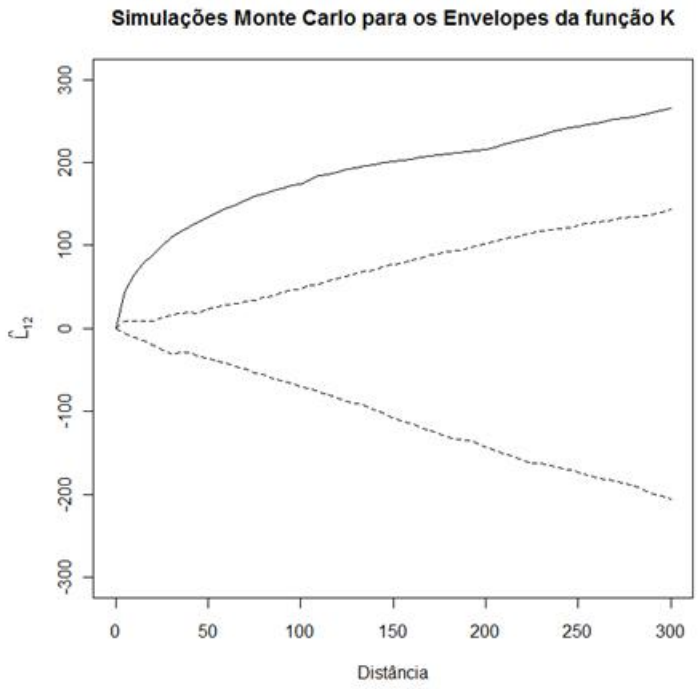

Figura 6 - Função K cruzada com envelopes de simulação após o deslocamento dos pontos coincidentes

A variável que identifica se o indivíduo é trabalhador terceirizado ou não é a única que de fato não é significativa no modelo inicial proposto. Pode-se também perceber que "anos de escolaridade"e "doméstico"também não são ao nível de 5% de significância.

A Tabela (7) apresenta o modelo selecionado por critério de seleção de variáveis segundo o critério AIC. Comparando com o modelo ajustado da Tabela (6) pode observar que as variáveis relacionadas ao risco referido (riscrf) e ao fato de ser trabalhador terceirizado ou não (tercei) não foram incluídas nesse ajuste.

A tabela (8) mostra estimativas para a modelagem GAM, na qual a componente espacial é levada em conta e inserida no processo de ajuste do modelo.

Com a inclusão da componente espacial a variável "anos de escolaridade"tornou-se mais não significante no ajuste do modelo. Outras variáveis que se tornaram não significativas foram: "Terceirizado"e "Doméstico".

A modelagem MAG ajustada para os dados em que os pontos coincidentes foram deslocados é mostrada na Tabela (9) a seguir.

Tabela 7 - Valores estimados para o MLG obtido com o critério de seleção de variável por AIC

Coeficientes	Estimativa	Erro Padrão	Estatística de teste	Valor p
Intercepto	-0.06184	0.25283	-0.245	0.80676
idade	-0.02481	0.00428	-5.792	<0.0001
sexo	0.84117	0.12814	6.565	<0.0001
anos escolaridade	-0.06482	0.01415	-4.580	<0.0001
ctps	0.69163	0.09526	7.260	<0.0001
na rua	-0.30139	0.10025	-3.006	0.00264
doméstico	-0.38771	0.17731	-2.187	0.02878

Tabela 8 - Valores estimados para o MLG com componente espacial suavizada

Coeficientes	Estimativas	Erro padrão	Estatística de teste	Valor p
Intercepto	-0.850135	0.298440	-2.849	0.00439
idade	-0.020886	0.004608	-4.532	<0.0001
sexo	0.859685	0.139653	6.156	<0.0001
anos escolaridade	-0.018795	0.014498	-1.296	0.19484
risco referido	0.042866	0.018091	2.370	0.01781
ctps	0.637901	0.124560	5.121	<0.0001
na rua	-0.301450	0.118583	-2.542	0.01102
terceirizado	0.030489	0.143329	0.213	0.83155
doméstico	-0.233061	0.201125	-1.159	0.24654
	edf.			Valor p
s(X,Y)	14			0.0706

A tabela (10) mostra o modelo GAM final ajustado. Observa-se que anos de escolaridade, as variáveis indicadoras de ser trabalhador terceirizado e de ser trabalhador doméstico não foram significativas na explicação da variação do risco.

A partir daí, ajustou-se um modelo multinomial ordinal classificando os dados como controle, casos leves e casos graves. Para esta modelagem, o nível considerado basal foi o

Tabela 9 - Valores estimados para o MLG com componente espacial suavizada para os dados com pontos coincidentes deslocados

Coeficientes	Estimativas	Erro padrão	Estatística de teste	Valor p
Intercepto	-0.837476	0.302352	-2.770	0.00561
idade	-0.020638	0.004651	-4.437	<0.0001
sexo	0.865013	0.140899	6.139	<0.0001
anos escolaridade	-0.021051	0.014743	-1.428	0.15333
risco referido	0.040655	0.018253	2.227	0.02593
ctps	0.647394	0.125716	5.150	<0.0001
na rua	-0.302167	0.119617	-2.526	0.01153
terceirizado	0.051784	0.144526	0.358	0.72012
doméstico	-0.211268	0.203227	-1.040	0.29854
	edf.			Valor p
s(X,Y)	9.81			0.0612

Tabela 10 - Valores estimados para o MLG com componente espacial suavizada para os dados com pontos coincidentes deslocados

Coeficientes	Estimativas	Erro padrão	Estatística de teste	Valor p
Intercepto	-1.16795	0.21981	-5.313	<0.0001
idade	-0.01882	0.00439	-4.288	<0.0001
sexo	0.95020	0.12284	7.735	<0.0001
risco referido	0.04430	0.01809	2.449	0.01431
ctps	0.66117	0.10626	6.222	<0.0001
na rua	-0.30866	0.11114	-2.777	0.00548
	edf.			Valor p
s(X,Y)	9.63			0.0297

grupo controle. Apesar de pouco utilizada, muitos eventos possuem distribuição multinomial. Para verificar sua eficiência, uma análise que incorpora resposta multinomial é mostrada na Tabela (11). Todas as estimativas incluídas nos modelos foram significativas ao nível de 5%

de significância.

Tabela 11 - Estimativas de graus de liberdade efetivos (edf) para as funções semiparamétricas relacionadas ao espaço e de odds ratios para as covariáveis paramétricas nos modelos ajustados para acidentes de trabalho em Piracicaba

	edf	*odds ratios* estimadas			
	s(x,y)	Sexo	CTPS(sim)	Idade	Anos escolaridade
Casos × Controles	13.80				
	11.15	2.357	1.937	0.191	0.939
Casos leves × Controles	11.11				
	8.05	2.139	2.101	0.004	0.931
Casos graves × Controles	10.10				
	9.40	3.056	1.521	467.603	0.955

A breve análise aqui realizada indica que a componente espacial é significativa no ajuste, implicando que existe variação espacial significativa do risco de sofrer acidente de trabalho na região estudada.

4.2.1 Abordagem com modelos estruturados aditivamente

Agora, sob o ponto de vista da abordagem para modelos estruturados para o mesmo conjunto de dados com ajuste via INLA, com a diferença de que será utilizada a análise considerados dados de área, serão mostrados os resultados obtidos comparando-os com os resultados obtidos na abordagem anterior por modelagem de um processo pontual.

As áreas utilizadas foram os setores censitários da área urbana de Piracicaba, o que corresponde a 507 setores censitários. Abaixo são mostrados os mapas do número de casos, números de controle e a razão entre o número de casos e de controles.

Observa-se que os setores censitários da parte direita do mapa (7) são os que apresentam maiores números de casos e de controles, com uma pequena diferença para o mapa de controles que apresenta outras áreas com quantidades maiores. Tais mapas, ajudam de forma inicial a identificação de padrões de interesse para o risco de sofrer acidente de trabalho nas áreas estudadas.

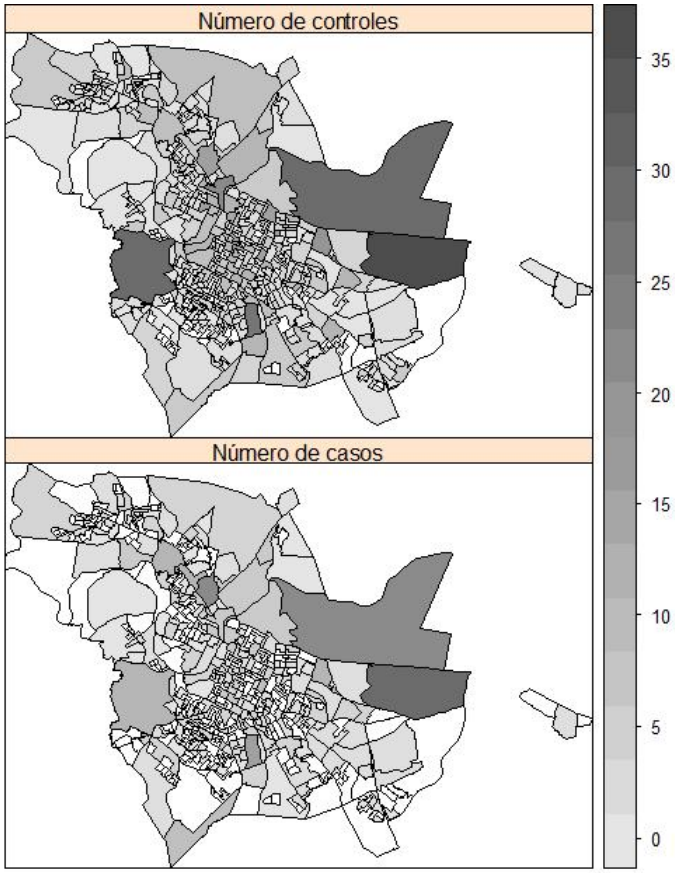

Figura 7 - Número de casos, número de controles por setor censitário

O mapa (8) da razão entre o número de casos e o número de controles não apresenta concentração em nenhuma área, o que é possível observar, é a existência de dois setores censitários na parte inferior esquerda do mapa com uma razão que apresenta valor alto, segundo a escala utilizada para a visualização do mapa (valores entre 0 e 5), conforme o resumo mostrado na Tabela (12).

Tabela 12 - Medidas resumo para a razão entre o número de casos e o número de controles segundo os setores censitários

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Áreas Vazias
0,05882	0,3782	0,6	0,8284	1,0	5,0	259

Vários modelos que levam em consideração a estrutura espacial de diferentes formas foram ajustados na ordem do mais simples para o mais complexo, os quais são mostrados

Figura 8 - Razão entre o número de casos e o número de controles por setor censitário

na tabela (13) abaixo e para comparação entre eles utilizou-se o Critério de informação da Deviance, o número estimado de parâmetros e a verossimilhança marginal.

Para o ajuste, utilizou-se como variável resposta o número de casos enquanto que, o número de controles foi utilizado como correção para a distribuição populacional da área estudada, cuja informação no processo de modelagem é feita atráves de um offset. Como notação para a variável resposta utilizou-se a letra Y e, φ para o efeito espacial estruturado e ϕ para o efeito espacial não estruturado.

Tabela 13 - Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal

Modelos	Preditor linear	DIC	NP	MV
1	$Y \sim 1$	1.017,98	1,004	0,0
2	$Y \sim \phi_i$	963,03	71,46	-502,16
3	$Y \sim \varphi_i$	967,15	50,65	-891,64
4	$Y \sim \phi_i + \varphi_i$	964,02	70,41	-886,94

O modelo 1 servirá de base de comparação e representa a variabilidade total contida nos dados. O modelo 2 considera, a priori, que o efeito espacial é não estruturado. O modelo 3 considera, a priori, que o efeito espacial é estruturado e, o modelo 4 considera que o efeito espacial é dividido em uma parte não estruturada e outra estruturada.

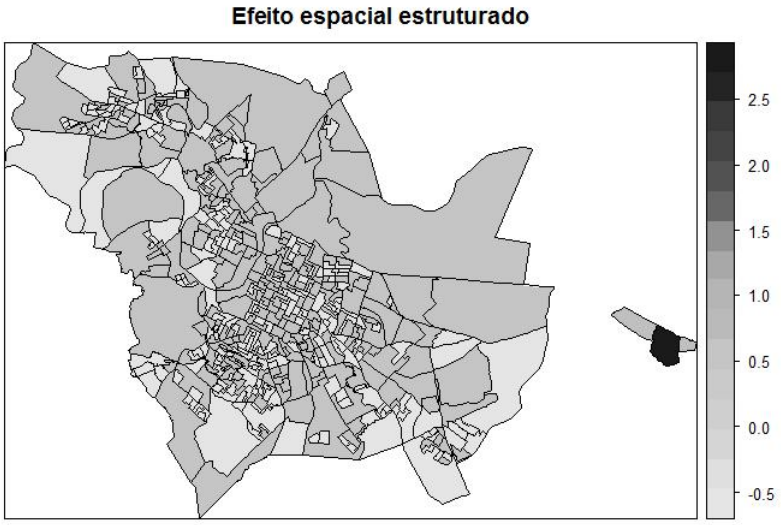

Figura 9 - Efeito espacial estruturado

Os resultados obtidos no processo de modelagem mostram que o modelo com menor DIC (Critério de informação da Deviance) é modelo 2, cujo efeito espacial não é estruturado a priori. Porém, considerando a existência de superdispersão nos dados, ou seja, parâmetro de variância maior que o parâmetro de média, o que sugere utilizar a distribuição binomial negativa com função de ligação logarítmica, obtém-se o resultado mostrado na Tabela 14.

Tabela 14 - Modelos ajustados, critério de informação da *Deviance*, número de parâmetros estimados e verossimilhança marginal considerando a superdispersão nos dados

Modelos	Preditor linear	DIC	NP	MV
1	$Y \sim 1$	986,97	1,675	-502,37
2	$Y \sim \phi_i$	987,63	2,062	-502,42
3	$Y \sim \varphi_i$	987,84	2,950	-887,27
4	$Y \sim \phi_i + \varphi_i$	987,98	3,060	-887,47

Todos os modelos apresentaram DIC muito próximos, sendo que o modelo que apresentou a menor verossimilhança marginal foi o modelo 4. O modelo 3 apresenta menor número de parâmetros e, o modelo 2 apresentou verossimilhança marginal muito próxima

da verossimilhança do modelo 1, que é a referência da variabilidade total. A princípio seria considerado o modelo 4, porém foi observado que o efeito espacial não estruturado é não significativo (Figura 12), portanto, o modelo a ser considerado será o modelo 3.

O histograma da Figura (10) mostra que o conjunto de dados está inflacionado de zeros, portanto, um ajuste pela distribuição binomial negativa é o mais adequado.

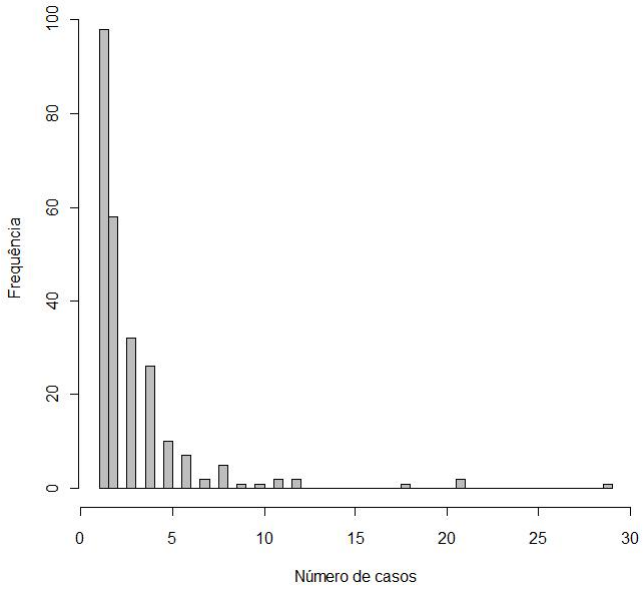

Figura 10 - Histograma do número de casos

A figura (11) mostra o ajuste do efeito espacial estruturado para as áreas analisadas. É observado que existe um pico no gráfico, implicando um efeito significativo do efeito espacial estruturado.

A figura (12) mostra o gráfico do ajuste do efeito espacial não estruturado. Observa-se que este efeito é não significativo e, que o intervalo de credibilidade é bastante largo.

O ajuste do modelo 3 (com efeito espacial estruturado) da tabela (14) para cada uma das covariáveis abaixo, considerando que a estrutura espacial já está representada no modelo é mostrada a seguir.

- Idade média
- Proporção de homens

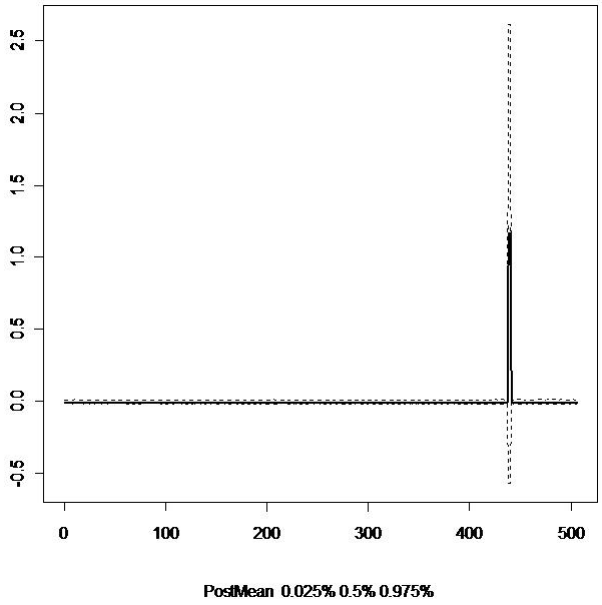

Figura 11 - Distribuição a posteriori - Modelo com efeito espacial estruturado

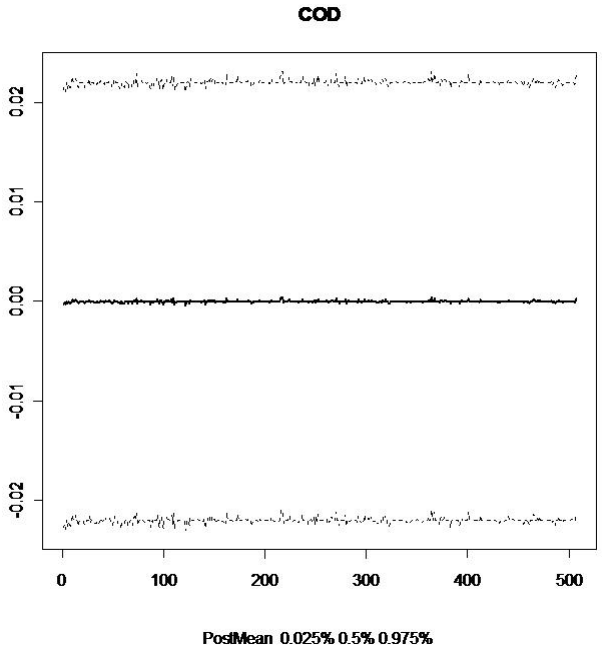

Figura 12 - Efeito espacial não estruturado

- Proporção de mulheres
- Proporção de trabalhadores com carteira assinada

- Proporção de trabalhadores sem carteira assinada
- Proporção de trabalhadores de rua
- Proporção de pessoas que não trabalham na rua
- Proporção de trabalhadores terceirizados
- Proporção de trabalhadores não terceirizados
- Proporção de trabalhadores domésticos
- Proporção de trabalhadores não domésticos

Um ajuste por MLG, considerando a distribuição binomial negativa para a variável resposta, as variáveis significativas foram: idade média, proporção de homens, proporção de trabalhadores com carteira assinada, proporção de trabalhadores de rua, proporção de trabalhadores terceirizados e proporção de trabalhadores domésticos, as quais foram utilizadas no ajuste por modelos estruturados aditivamente - abordagem INLA.

Controlando o efeito espacial é apresentado a seguir com as médias das distribuições a posteriori juntamente com os respectivos intervalos de credibilidade com nível nominal de 95%.

Tabela 15 - Ajuste do modelo com todas as covariáveis considerando efeito espacial estruturado e não estruturado

Covariável	Média a Posteriori	Intervalo de credibilidade
Idade média	-0,0375	(-0,0579; -0,0171)
Prop. cart. assinada	0,5084	(0,0439; 0,9683)
Prop. trab. de rua	-0,1921	(-0,6826; 0,3001)
Prop. terceirizados	0,2397	(-0,3116; 0,7823)
Prop. de domésticos	-0,7381	(-1,3861: -0,1037)

Segundo o resultado obtido na Tabela 15, as covariáveis que apresentaram significância, segundo os intervalos de credibilidade considerados com nível nominal de 95%,

foram: Idade média, proporção de pessoas com carteira de trabalho assinada e proporção de trabalhadores domésticos.

O modelo final no ajuste por modelos aditivos generalizados com componente espacial suavizada, obteve-se efeito significativo das seguintes covariáveis, além do efeito espacial:

- Idade
- Sexo
- Risco referido
- Carteira de trabalho
- Trabalha na rua

Na modelagem por INLA, optou-se por não utilizar a covariável "Risco referido" por se tratar de uma covariável com conteúdo apenas para aqueles trabalhadores que sofreram acidente de trabalho.

Os mapas das figuras (13), (14) e (15) mostram as distribuições das médias a posteriori dos efeitos das covariáveis consideradas significativas pelo ajuste com efeito espacial estruturado.

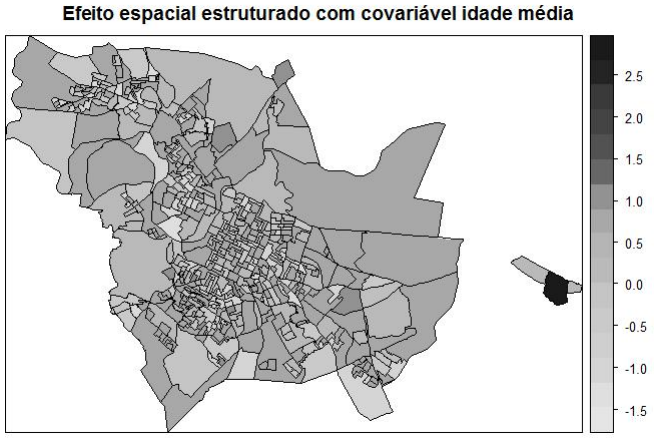

Figura 13 - Modelo com efeito espacial estruturado para a idade média

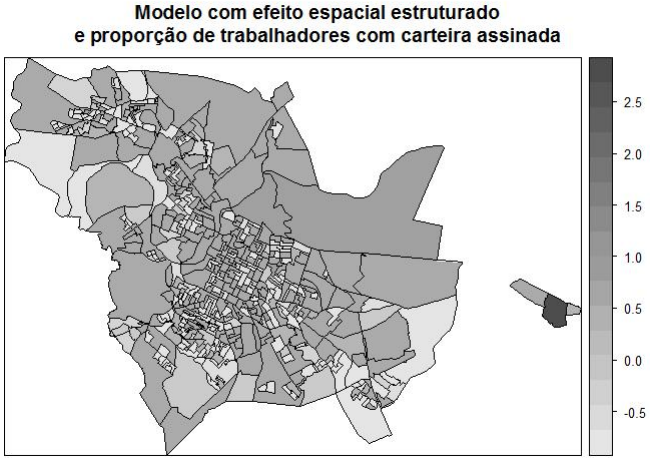

Figura 14 - Modelo com efeito espacial estruturado para a proporção de pessoas com carteira assinada

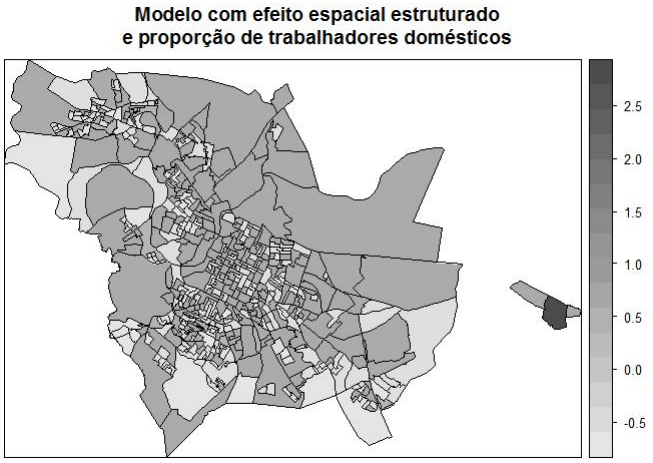

Figura 15 - Modelo com efeito espacial estruturado para a proporção de trabalhadores domésticos

A abordagem INLA identifica que, principalmente as áreas da borda direita da área estudada como sendo as de maiores médias a posteriori segundo as covariáveis consideradas para explicar a distribuição do número de trabalhadores que sofreram acidente decorrente do trabalho (casos).

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo mostrar metodologias que levam em consideração na modelagem de dados, o espaço na forma explícita ao longo do processo sob duas abordagens no intuito de compará-las. A primeira abordagem foi a modelagem por processo pontual considerando um processo pontual marcado, ou seja, dados com referências geográficas com informação adjacente às mesmas (caso ou controle). A segunda abordagem considerou a análise espacial por dados de áreas sob a abordagem de modelos estruturados aditivamente - INLA.

A metodologia aqui aplicada mostrou de modo inequívoco que o risco de sofrer acidente de trabalho entre os trabalhadores precarizados da área urbana de Piracicaba/SP varia de forma significativa no espaço e que, também, essa variação é modelada por covariáveis não espaciais.

A utilização da abordagem espacial em estudos caso-controle pretende mostrar que a consideração de forma explícita do espaço geográfico no processo de modelagem resulta em bons ajustes a dados que se apresentam georreferenciados e que possuem um padrão heterogêneo de ocupação do espaço.

Pretendeu-se obter como produto final mostrar que é possível identificar as áreas de sobre-risco, em qua a chance de sofrer acidente de trabalho é significantemente maior, mesmo quando se controla por covariáveis não espaciais. Vale ressaltar que, a metodologia que envolve análise espacial pode ser utilizada em diversas áreas de estudo e diversos tipos de dados, além de caso-controle.

A compreensão do comportamento do risco de sofrer acidente de trabalho e dos mecanismos que o governam são fundamentais para o controle da situação, com o intuito de evitar agravos de ocorrências.

O MAG, pouco utilizados na Epidemiologia, se mostrou uma ferramenta útil e abrangente na estimação da distribuição espacial do risco. Podem-se identificar áreas de maior prevalência do agravo, conjuntamente à análise espacial, analisar medidas de odds ratios para as outras covariáveis e, além disso, fazer comparações em diferentes âmbitos, como distribuições diferenciadas para a variável resposta e inclusão ou não de covariáveis.

Os modelos ajustados mostraram-se significativos na identificação das áreas de

baixo/alto risco. Foi importante a análise conjunta do espaço com as outras covariáveis, pois a inclusão destas modificou as estimativas. Conclusão: O MAG, pouco utilizados na Epidemiologia, se mostrou uma ferramenta útil e abrangente na estimação da distribuição espacial do risco.

A abordagem INLA utilizada tende a ser mais conservadora que a abordagem MAG, por isso, para que uma covariável seja considerada significativa, ela precisa trazer muito mais informação do que é exigido pela abordagem por MAG, onde não se tem distribuição a priori. É provável que, dentre outros fatores, essa característica tenha sido a principal causa da diferença de resultados encontrados entre as duas abordagens utilizadas.

A abordagem INLA é muito recente e, pode-se dizer que até inédita na aplicação em dados epidemiológicos. Tendo-se, portanto, um amplo caminho pela frente de exploração deste método, seja por modelos mais complexos ou outras variantes.

É possível, ainda, recorrer a outros procedimentos estatísticos que consideram o espaço no processo de análise de dados, os quais podem ser abordados em futuros trabalhos, tais como, por exemplo, a abordagem por análise de superfícies contínuas, o que reforça o que fora dito em parágrafos anteriores, ainda há muito o que se explorar em se tratando de análise espacial de dados e, principalmente sob a abordagem INLA (INTEGRATED NESTED LAPLACE APPROXIMATIONS).

REFERÊNCIAS

ABREU, M.N.S; SIQUEIRA, A.L.; CAIAFFA, W.T. Regressão logística ordinal em estudos epidemiológicos. **Revista de Saúde Pública**, São Paulo, v. 43, n. 1, p. 183-94. 2009.

ASSUNÇÃO, R. **Estatística espacial com aplicações em epidemiologia, economia e sociologia**. São Carlos: UFSCar, 2001.

BAILEY, T.C.; GATTREL, T.C. **Interactive spatial data analysis**. London: Prentice Hall, 1995.

BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesian image restoration with two applications in spatial statistics. **Annals of Institute of Statistical Mathematics**, Tokyo, v.43, n. 1, p.1-59, 1991.

BIVAND, R.S.; PEBESMA, E.J.; GÓMEZ-RUBIO, V. **Applied spatial data analysis with R**. New York: Springer, 2008.

BONAT, W.H. **Aplicações de inferência Bayesiana aproximada para modelos Gaussianos latentes espaço temporais**. 2010. 80p. Dissertação (Mestrado em Métodos Numéricos em Engenharia) - Centro de Estudos de Engenharia Civil Professor Inaldo Ayres Vieira, Universidade Federal do Paraná, Curitiba, 2010.

CÂMARA, G.; MONTEIRO, A. M.; DAVIS, C. **Introdução à ciência da geoinformação**. São José dos Campos: INPE, 2003.

CARVALHO, M.S.; SOUZA-SANTOS, R. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. **Caderno de Saúde Pública**, Rio de Janeiro, v. 21, n. 2, p. 361-378, 2005.

CONCEIÇÃO, G. M. S.; SALDIVA, P. H. N.; SINGER, J. M. Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo. **Revista Brasileira de Epidemiologia**, São Paulo, v. 4, n.3, p. 206-219, 2001.

CRUZ, C.M.; BARROS, R.S. **Análise do padrão de distribuião espacial do índice de equidade sócio-econômica no município do Rio de Janeiro**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais. Centro de Estudos de Desigualdades Sócio-Territoriais, 2000. (Relatório Técnico)

DeGROOT, M.H. **Probability and statistics**. 2nd. ed. Pittsburgh: Addison-Wesley Publishing Company, 1989. 723p.

DIGGLE, P.J. **Statistical analysis of spatial point patterns**. 2nd. ed. London: Arnold, 2003. 159p.

DIGGLE, P. J.; RIBEIRO JR., P. J. **Model-based geostatistics**. New York: Springer, 2007. 228p.

DRUCK, S.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A. M. V. **Análise espacial de dados geográficos**. Brasília: EMBRAPA, 2004.

EIDSVIK, J.; MARTINO, S.; RUE, H. Approximate Bayesian inference in spatial generalized linear mixed models. **Scandinavian Journal of Statistics**, Stockholm, v. 36, p. 1-22, 2009.

FAHRMEIR, L.; TUTZ, G. **Multivariate statistical modelling based on generalized linear Models**. 2nd ed. Berlim: Springer-Verlag, 2001.

FIGUEIRA, C.V. **Modelos de regressão logística**. 2006. 138 p. Dissertação (Mestrado em Matemática) - Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

GAMERMAN, D. **Markov chain Monte Carlo**: stochastic simulation for Bayesian inference. Texts in Statistical Sciences. London: Chapman and Hall, 1997.

HASTIE, T.J.; TIBISHIRANI, R.J.**Generalized additive models**. London: Longman, 1990.

HOSMER, D.W.; LEMESHOW, S. **Applied logistic regression**. New York: Wiley, 1989.

KELSALL, J. E.; DIGGLE, P. J. Spatial variation in risk of disease: a nonparametric binary regression approach. **Applied Statistics**. Lancaster, v.47, 559-573, 1998.

MCCULLAGH P. Regression models for ordinal data. **Journal of the Royal Statistical Society Series B**, Oxford, v. 42, n. 2, p. 109-42, 1980.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society Series A**, Oxford, v.135, p.370-384, 1972.

OLINDA, R.A. **Métodos para análise de independência entre marcas e pontos em processos pontuais marcados**. 2008. 76 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, 2008.

PAULA, G.A. **Modelos de regressão com apoio computacional**. São Paulo: IME-USP, 2004. 245p.

PICHARSKI, G.L.; RIBEIRO Jr., P.J.; SHIMAKURA, S.E.; CARVALHO, M.L.; MOYSÉS, S.J.; MOYSÉS, S.T. Metodologias para análise de dados pontuais: casos de trauma dentário em Curitiba. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA 54., SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA 13, 2009, São Carlos. **Anais ...** São Carlos: EdUSFSCar. 2009.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.r-project.org>. Acesso em: 10 de dezembro de 2009.

RESENDE, M.D.V.; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com respostas binomiais. **Revista de Matemática e Estatística**, São Paulo, v. 20, p. 39-65, 2002.

RIPLEY, B.D. Modelling spatial patterns. **Journal of the Royal Statistical Society**, Oxford, v.39, n. 2, p. 172-212, Jun. 1977.

ROBERT, C.P.; CASELLA,G. **Monte Carlo statistical methods**. New York: Springer-Verlag, 1999.

RUE, H.; HELD, L. **Gaussian Markov random fields:** theory and applications. London: Chapman & Hall, 2005.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. **Journal of the Royal Statistical Society Series B**, Oxford, v. 71, p. 319-392, 2009.

SHIMAKURA, Silvia E.; CARVALHO, Marilia Sá; AERTS, Denise R. G. C.; FLORES, Rui. Distribuição espacial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. **Caderno de Saúde Pública**, São Paulo, v.17, n.5, p. 1251-1261. 2001.

SPIEGELHALTER, D.J.; BEST, N.G.; CARLIN, B.P.; LINDE, A.V.D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society Series B**, Oxford, v.64, p. 583-639, 2001.

WAND, M.; JONES, M.C. **Kernel smoothing**. London: Chapman and Hall. 1995.

APÊNDICE

BIBLIOTECAS NECESSÁRIAS PARA O PROCESSAMENTO

```
require(spdep)
require(mgcv)
require(splancs)
require(spgam)
require(maptools)
require(INLA)

###################### # LEITURA DOS DADOS # ######################

dados=read.table('dados.csv',header=T,sep='\t',dec='.')
dados1<-dados
coordinates(dados1)=~x+y

mm <- sapply(1:(dim(dados1)[1]),
function(i)spDistsN1(coordinates(dados1),coordinates(dados1)[i,]))

mm <- sapply(1:(dim(dados.sp)[1]),
function(i)spDistsN1(coordinates(dados.sp),coordinates(dados.sp)[i,]))

mm1 <- cbind(expand.grid(1:(dim(dados1)[1]),1:(dim(dados1)[1])),
as.vector(mm))

mm1 <- cbind(expand.grid(1:(dim(dados.sp)[1]),
1:(dim(dados.sp)[1])),as.vector(mm))

mm2 <- mm1[mm1[,1]<mm1[,2]&mm1[,3]==0,] mm3 <-
sort(unique(as.vector(as.matrix(mm2[,1:2]))))

x1=dados1$x y1=dados1$y

set.seed(100)
x1[mm3]=sapply(x1[mm3],function(x)runif(1,x-raio,x+raio))
set.seed(100)
y1[mm3]=sapply(y1[mm3],function(x)runif(1,x-raio,x+raio))

dados.sp <- dados1
dados.sp$x <- x1
dados.sp$y <- y1
write.csv(dados.sp,"dados2.csv")

dados2 = read.csv("dados2.csv",head=T)
dados3 <- dados2
coordinates(dados3) = ~x+y

##### Desenha o contorno da área estudada ####
```

```
set.seed(100)
oldnew
<-cbind(dados$x,dados$y,dados2$x,dados2$y)[sample(mm3,200),]
plot(rbind(oldnew[,1:2],oldnew[,3:4]),ty="p",cex=0.5) i=1
sapply(1:(dim(oldnew)[1]),function(i)segments(oldnew[i,1],
oldnew[i,2],oldnew[i,3],oldnew[i,4]))
sapply(1:(dim(oldnew)[1]),function(i)arrows(oldnew[i,1],
oldnew[i,2],oldnew[i,3],oldnew[i,4],lenght=0.1))

cont <- as.matrix(read.table("poly1.txt"))

piracicaba <-readShapePoly("piracicaba")

################## # GRUPO CONTROLE # ##################

# Não caso para dados originais
nao.caso <-
as.points(as.matrix(dados1[dados1$tipfch == 0,c("x","y")]))

# Não caso para dados deslocados

nao.caso <- as.points(as.matrix(dados3[dados3$tipfch ==
0,c("x.1","y.1")]))

# ############## # GRUPO CASO # ############## # #

Caso para dados originais \\

caso <- as.points(as.matrix(dados1[dados1$tipfch == 1,c("x","y")]))

# Caso para dados deslocados

caso <- as.points(as.matrix(dados2[dados2$tipfch ==
1,c("x.1","y.1")]))

##### KERNEL E RAZÃO DE KERNEL #####

par(mfrow=c(1,3))

# kernel.c <- kernel2d(caso,cont,h0=1000,nx=400,ny=400)
image(kernel.c,col=rev(gray.colors(12))) title("Kernel dos casos") #

kernel.nc <- kernel2d(nao.caso,cont,h0=1000,nx=400,ny=400)

image(kernel.nc,col=rev(gray.colors(12))) title("Kernel dos
controles") #

kernel.raz <-
```

```
kernrat(caso,nao.caso,cont,h1=3000,h2=3000,nx=400,ny=400)
image(kernel.raz,col=rev(gray.colors(12))) #with(kernel.nc,
legend.krige(c(685000,686000),c(7165000,7195000),vert=T,z,
col=rev(gray.colors(12)),cex=1.3))
title("Kernel razão")

#### FUNÇÃO K CRUZADA ####

contorno <- as.data.frame(cont) coordinates(contorno)=~V1+V2

a<-seq(0,300,5)
kcruzada <- k12hat(caso,nao.caso,cont,a)
plot(a,
sqrt(kcruzada/pi) - a, xlab="Distância",ylab=expression(hat(L)[12]),
ylim=c(-300,300), type="l",main="Simulações Monte Carlo para os
Envelopes da função K")
env.ok <-
Kenv.tor(caso,nao.caso,cont,nsim=29,s= a) lines(a,
sqrt(env.ok$upper/pi)-a, lty=2) lines(a, sqrt(env.ok$lower/pi)-a,
lty=2)

#### GLM ####

dados1[dados1== "-"] <- NA
dados1$tipaci <-
as.factor(as.character(dados1$tipaci))

# transformando sexpes em fator

dados1$sexpes <- as.factor(as.character(dados1$sexpes))

# transformando riscrf em fator

dados1$riscrf <- as.factor(as.character(dados1$riscrf))

#fit1<-glm(tipfch~idade,data=dados,family=binomial("logit"))

#fit2 <-glm(tipfch~idade+sexpes,data=dados,family=binomial("logit"))

#fit3
<-glm(tipfch~idade+sexpes+anoesc,data=dados,family=binomial("logit"))

#fit4
<-glm(tipfch~idade+sexpes+anoesc+riscrf,data=dados,
family=binomial("logit"))

#fit5
<-glm(tipfch~idade+sexpes+anoesc+riscrf+ctps,data=dados,
```

```
family=binomial("logit"))

#fit6
<-glm(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua,data=dados,
family=binomial("logit"))

#fit7
<-glm(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei,
data=dados,family=binomial("logit"))

#fit8
<-glm(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei+domest,
data=dados.sp,family=binomial("logit"))

# fit
<-step(glm(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei+domest,
data=dados1,family=binomial("logit")))
summary(fit,cor=F)

##### GAM #####

dados2 = read.csv("dados2.csv",head=T)
dados3 <- dados2 #
dados3[dados3== "-"] <- NA
dados3$tipaci <-
as.factor(as.character(dados3$tipaci))

# transformando sexpes em fator

dados3$sexpes <- as.factor(as.character(dados3$sexpes))

# transformando riscrf em fator

dados3$riscrf <- as.factor(as.character(dados3$riscrf))

fit9 <-
gam(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei+
domest+s(x,y,bs="tp"),data=dados1,family=binomial("logit"))

fit9 <-
gam(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei+
domest+s(x,y,bs="tp"),data=dados3,family=binomial("logit"))
summary(fit9,cor=F)

# # Exclusão de tercei

fit10 <-
gam(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+domest+
```

```
s(x,y,bs="tp"),data=dados3,family=binomial("logit"))

summary(fit10,cor=F)

# # Exclusão de tercei + domest

fit11 <-
gam(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+
s(x,y,bs="tp"),data=dados3,family=binomial("logit"))

summary(fit11,cor=F)

# # Exclusão de  domest

fit12 <-
gam(tipfch~idade+sexpes+anoesc+riscrf+ctps+narua+tercei+
s(x,y,bs="tp"),data=dados3,family=binomial("logit"))
summary(fit12,cor=F) # # Exclusão de tercei+domest+anoesc

fit13 <-
gam(tipfch~idade+sexpes+riscrf+ctps+narua+s(x,y,bs="tp"),
data=dados3,family=binomial("logit"))
summary(fit13,cor=F)

# # Comandos para uma modelagem ordinal:

dados<-"C:/Users/Marcelo
Tavares/Documents/ESALQ/Dissertacao/dados/AcdeTrab.RData"
load(dados)
table(dados$graaci2)
library(VGAM) # Politômicos # GLM:
poli_glm<-vglm(graaci2~idade+sexpes+anoesc+riscrf,
family=multinomial, dados) # GAM:
#fit14<-vglm(graaci2~s(x)+s(y)+idade+sexpes+anoesc+riscrf,
multinomial, dados)
fit14<-vglm(graaci2~idade+sexpes+anoesc+riscrf+ctps+narua
+tercei+domest+s(x)+s(y),
multinomial, dados) summary(poli_glm) summary(fit14, cor=F)

#### ANÁLISE VIA INLA ####

setwd('C:/Users/Marcelo Tavares/Documents/ESALQ/Dissertacao/dados')

#### LEITURA DOS DADOS ####
dados=read.table('dados.csv',header=T,sep='\t',dec='.')

# ######### Bases limpas
```

```
dados.new <- dados[,c(4,7,8,9,10,11,12,13,14,15,5,6)]

teste <- data.frame(coordinates(dados.new))
coordinates(teste) =~x+y
coordinates(dados.new) = ~x+y
conta.casos <- overlay(mapa,teste)

dados.na = na.exclude(dados.new)

base.setor <- data.frame(Setores = levels(mapa$ID_),N.Casos = NA,
N.controle = NA, N.total = NA, Idade.m = NA, Masc = NA, Fem= NA,
CTPS.S = NA, CTPS.N = NA, RUA.S = NA, RUA.N = NA, TER.S = NA, TER.N
= NA, DOM.S = NA, DOM.N = NA)

dados.na[which(dados.na$ID == base.setor$Setores[1]),]

funcao.calculo <- function(dados.setor){ n <- dim(dados.setor)[1]
N.controle = as.numeric(table(dados.setor$tipfch)[1]) N.casos =
as.numeric(table(dados.setor$tipfch)[2]) N.total = N.controle +
N.casos Idade.m = mean(dados.setor$idade) Fem =
as.numeric(table(dados.setor$sexpes)[2])/n Masc=
as.numeric(table(dados.setor$sexpes)[1])/n CTPS.N =
as.numeric(table(dados.setor$ctps)[1])/n CTPS.S =
as.numeric(table(dados.setor$ctps)[2])/n RUA.N =
as.numeric(table(dados.setor$narua)[1])/n RUA.S =
as.numeric(table(dados.setor$narua)[2])/n

TER.N = as.numeric(table(dados.setor$tercei)[1])/n TER.S =
as.numeric(table(dados.setor$tercei)[2])/n

DOM.N = as.numeric(table(dados.setor$domest)[1])/n DOM.S =
as.numeric(table(dados.setor$domest)[2])/n

saida = round(data.frame(N.Casos = N.casos, N.controle = N.controle,
N.total = N.total, Idade.m = Idade.m, Masc = Masc, Fem= Fem, CTPS.S
= CTPS.S, CTPS.N = CTPS.N, RUA.S = RUA.S, RUA.N = RUA.N, TER.S =
TER.S, TER.N = TER.N, DOM.S = DOM.S, DOM.N = DOM.N),2)
return(saida)}

for(i in 1:507){ base.setor[i,][2:15] <-
funcao.calculo(dados.na[which(dados.na$ID ==
base.setor$Setores[i]),])}

require(spdep)
mat.viz = poly2nb(mapa,row.names=mapa$region.id)

base.setor$COD <- 1:507
```

```
#cat(507,file="teste.graph",append=TRUE,fill=TRUE,sep=" ") #for(i in
1:507){ #cat(c(i-1, length(mat.viz[[i]]),mat.viz[[i]]-1), sep=" ",
append=TRUE,fill=TRUE,file="teste.graph")}


formu <- N.Casos ~ f(COD,model="besag",graph.file="teste.graph")
modelo = inla(formu,family="poisson",data=base.setor,E=N.controle)
summary(modelo) plot(modelo)

# criando variável que caracteria o efeito aleatório

dados$COD.iid <- dados$COD

#### MODELAGEM VIA INLA ####

#formu <- N.Casos ~ f(COD,model="besag",graph.file="teste.graph")
#modelo = inla(formu,family="poisson",data=dados,E=N.controle)
#summary(modelo)
plot(modelo)

# # modelo com apenas o intercepto

formu1 <- N.Casos ~ 1 modelo =
inla(formu1,family="poisson",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE)) summary(modelo)
plot(modelo)

# # modelo com efeito espacial estruturado

formu <- N.Casos ~ f(COD,model="besag",graph.file="teste.graph")
modelo = inla(formu,family="poisson",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE)) summary(modelo)
plot(modelo)

# # modelo com efeito espacial não estruturado

formu <- N.Casos ~ f(COD,model="iid") modelo =
inla(formu,family="poisson",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE)) summary(modelo)
plot(modelo)

# # modelo com efeito espacial estruturado e não estruturado
formu.p.4 <- N.Casos ~ f(COD,model="besag",graph.file="teste.graph")
+ f(COD.iid,model="iid") modelo.p.4 =
inla(formu.p.4,family="poisson",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.p.4) plot(modelo)
```

```
# # # modelo com idade, sexo, ctps, narua e efeito espacial
estruturado formu5 <- N.Casos~Idade.m + mulher + homem + CTPS.S +
CTPS.N + RUA.S + RUA.N +
f(COD,model="besag",graph.file="teste.graph") modelo5 =
inla(formu5,family="poisson",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo5) plot(modelo5) #

## INLA com a distribuição binomial negativa##

histograma do número de casos hist(dados$N.Casos, breaks=50,
col="gray", main=NULL, xlab="Número de casos", ylab="Frequência")

# # modelo com apenas o intercepto

formu.bn.1 <- N.Casos ~ 1 modelo.bn.1 =
inla(formu.bn.1,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.1) plot(modelo)

# # modelo com efeito espacial estruturado

formu.bn.2 <- N.Casos ~
f(COD,model="besag",graph.file="teste.graph") modelo.bn.2 =
inla(formu.bn.2,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.2) plot(modelo)

mapa.new@data <- data.frame(mapa.new@data,
modelo.bn.2$summary.fitted.values[1])

# Mapa do efeito espacial estruturado

spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80',
'gray70','gray30','gray10'))(20),
main='Efeito espacial estruturado')

# # modelo com efeito espacial não estruturado

formu <- N.Casos ~ f(COD,model="iid")

modelo = inla(formu,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo)
plot(modelo)

# modelo com efeito espacial estruturado e não estruturado
```

```
formu <- N.Casos ~ f(COD, model="besag", graph.file="teste.graph") +
f(COD.iid,model="iid")
modelo
=inla(formu,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo)
plot(modelo)

## MAPAS
require(spdep)
mapa = readShapePoly("C:/Users/Marcelo
Tavares/Documents/ESALQ/Dissertacao/dados/shape/mapa_setores_pol",
proj4string=CRS("+proj=utm +zone=23 +units=m +south"))

# modelo com efeito espacial estruturado
formu.bn.2 <- N.Casos ~
f(COD,model="besag",graph.file="teste.graph")
modelo.bn.2 =
inla(formu.bn.2,family="nbinomial",data=dados.new2,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.2)

# limpando o mapa
mapa.new <- mapa mapa.new@data <-
mapa.new@data[,c(1,2,3,4,5,6,7,8,9)]

# junção da média a posteriori do efeito espacial estruturado
mapa.new2@data <- data.frame(mapa.new2@data,
modelo.bn.2$summary.fitted.values[1])

#################################################################
modelo com efeito espacial estruturado e todas as covariáveis
#################################################################

library(MASS)
m1 <-
glm.nb(N.Casos~offset(N.controle)+Idade.m+homem+mulher+
CTPS.S+CTPS.N+RUA.S+RUA.N+
TER.S+TER.N+DOM.S+DOM.N,data=dados,etastart=0,mustart=0)

# modelo para a idade média
formu.bn.idade.m <- N.Casos~Idade.m +
f(COD,model="besag",graph.file="teste.graph")
modelo.bn.idade.m =
inla(formu.bn.idade.m,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.idade.m)
```

```
plot(modelo.bn.idade.m)

# inserindo a idade média nos dados
mapa.new@data <-
data.frame(mapa.new@data,
modelo.bn.idade.m$summary.fitted.values[1])

# Mapa do efeito espacial estruturado com a presença da covariável
idade média
spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80',
'gray70','gray30','gray10'))(20),
main='Efeito espacial estruturado com covariável idade média')

#limpando o banco dados : excluindo a variavel MAsc e FEM -
dados$MAsc <- NULL dados$FEM <- NULL
write.table(dados, "dados.txt")

# Modelo com efeito espacial estruturado e covariável proporção de
pessoas com carteira assinada
formu.bn.c.a <- N.Casos~ CTPS.S +
f(COD,model="besag",graph.file="teste.graph")
modelo.bn.c.a =
inla(formu.bn.c.a,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.c.a)
plot(modelo.bn.c.a)

# unindo as informações
mapa.new@data <- data.frame(mapa.new@data,
modelo.bn.c.a$summary.fitted.values[1])

# Mapa de número de efeito da covariável proporção de trabalhadores
com carteira assinada + efeito espacial estruturado

spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80',
'gray70','gray60','gray50','gray40','gray30'))(30),
main='Modelo com efeito espacial estruturado\ne proporção de
trabalhadores com carteira assinada')

# Modelo com efeito espacial estruturado e covariável proporção de
trabalhadores de rua

formu.bn.r.s <- N.Casos~ RUA.S +
f(COD,model="besag",graph.file="teste.graph") modelo.bn.r.s =
inla(formu.bn.r.s,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
```

```
summary(modelo.bn.r.s) plot(modelo.bn.r.s)

# unindo as informações

mapa.new@data <- data.frame(mapa.new@data,
modelo.bn.r.s$summary.fitted.values[1])

# Mapa de número de efeito da covariável proporção de trabalhadores
de rua + efeito espacial estruturado

spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
main='Modelo com efeito espacial estruturado\ne proporção de
trabalhadores de rua')

#================================================================ #
Modelo com efeito espacial estruturado e covariável proporção de
trabalhadores terceirizados

formu.bn.terc <- N.Casos~ TER.S +
f(COD,model="besag",graph.file="teste.graph") modelo.bn.terc =
inla(formu.bn.terc,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
summary(modelo.bn.terc) plot(modelo.bn.terc)

# unindo as informações mapa.new@data <- data.frame(mapa.new@data,
modelo.bn.terc$summary.fitted.values[1])

# Mapa de número de efeito da covariável proporção de trabalhadores
terceirizados + efeito espacial estruturado

spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
main='Modelo com efeito espacial estruturado\ne proporção de
trabalhadores terceirizados')
#================================================================

# Modelo com efeito espacial estruturado e covariável proporção de
trabalhadores domésticos

formu.bn.dom <- N.Casos~ DOM.S +
f(COD,model="besag",graph.file="teste.graph")

modelo.bn.dom =
inla(formu.bn.dom,family="nbinomial",data=dados,E=N.controle,
control.compute=list(dic=TRUE, cpo=TRUE, mlik=TRUE))
```

```
summary(modelo.bn.dom) plot(modelo.bn.terc)

# unindo as informações mapa.new@data <- data.frame(mapa.new@data,
modelo.bn.dom$summary.fitted.values[1])

# Mapa de número de efeito da covariável proporção de trabalhadores
de rua + efeito espacial estruturado

spplot(mapa.new, "mean",
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
main='Modelo com efeito espacial estruturado\ne proporção de
trabalhadores domésticos')
#============================================================

# Mapa de número de casos

spplot(mapa.new, "N.Casos",
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
main='Número de Casos')

# Mapa de número de controles

spplot(mapa.new, "N.controle",
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
main='Número de Controles')

# criando uma variavel para a razão entre casos e controles
dados$razao <- dados$N.Casos/dados$N.controle

# Mapa para o número de casos, número de controles juntos
spplot(mapa.new, c("N.Casos","N.controle"),
col.regions=colorRampPalette(c('gray90','gray80','gray70',
'gray60','gray50','gray40','gray30'))(30),
names.attr = c("Número de casos", "Número de controles"))

# Mapa para a razão entre o número de casos e o número de controles
spplot(mapa.new, "razao",
col.regions=colorRampPalette(c('gray90','gray80','gray60',
'gray40','gray30'))(30))
```