UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

FATORES DE RISCO E DISTRIBUIÇÃO ESPACIAL DO RISCO PARA A MORTALIDADE INFANTIL EM CURITIBA - 2004

CURITIBA-PR
NOVEMBRO 2006

VALÉRIA VERÍSSIMO FRANCO DE SOUSA
FELIPE EMANOEL BARLETTA MENDES

FATORES DE RISCO E DISTRIBUIÇÃO ESPACIAL DO RISCO PARA A MORTALIDADE INFANTIL EM CURITIBA - 2004

Trabalho de Conclusão de Curso apresentado à banca examinadora como requisito para graduação no curso de Estatística, tendo como orientadora de conteúdo a Professora Silvia Emiko Shimakura.

CURITIBA-PR
NOVEMBRO 2006

EQUIPE TÉCNICA

Departamento de Estatística
Setor de Ciências Exatas

Aluna de Graduação
Valéria Veríssimo Franco de Sousa

Aluno de Graduação
Felipe Emanoel Barletta Mendes

Orientadora
Prof. Silvia Emiko Shimakura, Dra.

AGRADECIMENTOS

A Deus por permitir que pudéssemos concluir este trabalho.

Ao meu esposo Ricardo que me deu força para que chegasse até o fim (Valéria).

Aos nossos pais que nos ensinaram que o conhecimento é o único bem que podemos dizer que realmente nos pertence.

A empresa SHIFT TECHNOLOGY por possibilitar a *linkagem* dos bancos de dados.

Ao IPPUC, e principalmente a Lourival Peyerl que nos deu grande apoio, e a Margareth Rose Kolb (Meg) por georeferenciar o banco de dados, tornando assim possível a realização este trabalho.

E especialmente a Professora Silvia Emiko Shimakura, pela colaboração e orientação que nos auxiliaram na conclusão deste estudo.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

Um desafio para os profissionais da área de epidemiologia é mensurar os efeitos espaciais que atuam sobre eventos de saúde pública. Neste trabalho é analisada a distribuição espacial da mortalidade infantil em Curitiba, Paraná, ocorridas em 2004, tendo como componente para controle da heterogeneidade dos casos ao longo do espaço urbano estudado, uma amostra de nascidos vivos que foi extraída do Sistema de Informações de nascidos vivos da mesma cidade. Desta forma foi utilizado um modelo semiparamétrico para a modelagem de uma medida de risco que pode variar suavemente ao longo do espaço. A abordagem é realizada através de um Modelo Aditivo Generalizado que possui termos paramétricos e termos não-paramétricos, aproxima funções suaves de acordo com os dados analisados, assim estimando a variação continuamente ao longo do espaço. Também se determina a identificação de áreas com maior ou menor risco, através da plotagem no mapa de Curitiba da superfície de risco estimada. A aplicação de todos estes aspectos e metodologia de análise mostrou uma variação espacial significativa para a componente pós-neonatal, o que não ocorreu para a componente neonatal.

ABTRACT

A challenge for professionals operating in the field of epidemiology id to measure the spatial effects which influence public health events. This work analyses the spatial distribution of infant mortality in Curitiba, Paraná, which occurred in 2004, and has as a heterogeneous control component of cases alongside the urban space studied, a live-birth sample which was taken from the information System of live-births of the same city. In this way, a semi-parametric model was used for the modelling of a risk medium which can mildly vary alongside space. The approach is carried out by a Generalized Additive Model which possesses parametric terms and non-parametic terms, drawing close to gentle functions in agreement with the analysed data, thus estimating the variation continually alongside space. Also, the identification of areas is determined with greater or lesser risk, by plotting on the map of Curitiba the surface of estimated risk. The application of all of these aspects and analysis methodology showed a significant spatial variation for the post-neonatal component, which doesn't occur for the neonatal component.

1. INTRODUÇÃO

Na área epidemiológica a mortalidade infantil é com freqüência usada como indicador dos mais sensíveis da condição de saúde de uma região (SOUZA et al, 1993), além disso, os óbitos infantis estão mais associados a fatores sociais do que os óbitos ocorridos na idade adulta. Uma sub-divisão ainda mais apurada considera a mortalidade infantil dividida em dois componentes: mortalidade neonatal (óbitos de crianças com menos de 28 dias de idade) e mortalidade pós-neonatal (óbitos ocorridos entre 28 dias a menos de 1 ano). É sabido que a mortalidade neonatal está habitualmente associada a anomalias congênitas da criança ou a complicações da gravidez ou/e do parto, e a mortalidade pós-neonatal está mais associada às condições de vida, deficiências sanitárias e causas externas e, por isso mesmo, mais permeável a intervenções que permitam melhorar este indicador.

As taxas de mortalidade infantil no Brasil costumam ser mais fidedigna quando associadas às informações demográficas como é descrito por Kozu (2006). Conhecer o perfil da mortalidade infantil é imprescindível para traçar planos mais eficientes em relação ao seu controle. Para isso tornar-se realidade é importante um apanhado mais sucinto em relação ao problema, então se deve levar em consideração as taxas de mortalidade neonatal e pós-neonatal e onde elas ocorreram e também alguns fatores que freqüentemente estão associados à Mortalidade Infantil, como: status financeiro, raça, escolaridade dos pais e localização geográfica da residência.

Devido aos grandes avanços computacionais, tanto em relação à capacidade de armazenamento de grandes massas de dados quanto em relação ao processamento destes, estas progressões vêm viabilizando cada vez mais trabalhos híbridos entre SIG (Sistemas de Informações Geográficas) e técnicas mais elaboradas em Análise de Dados Espaciais. Os SIG, que por sua vez também vem sofrendo grande desenvolvimento e conseqüentemente cria grandes bases de dados Georreferenciadas de qualidade, disponibilizam as informações demográficas tão úteis para o monitoramento da mortalidade infantil. Técnicas estatísticas sofisticadas para este tipo de análise, como é o caso do Modelo Aditivo Generalizado, também tiveram avanços interessantes, trazendo assim perspectivas inovadoras.

Estes aspectos permitem estudos mais precisos na área da epidemiologia e saúde pública, já que a introdução de informações geográficas permite o gerenciamento das influências sócio-culturais e sócio-econômicas que estão relacionadas diretamente com meio em análise. Portanto os estudos não estritamente individuais (Carvalho et al,2005) que levam em consideração os efeitos espaciais e efeitos em nível de grupo mostram-se mais adequados para problemas de natureza etiológica e detecção de padrões de mortalidade. O mapeamento de perfis de risco vem sendo instrumento básico no campo da saúde pública e, em anos recentes, segundo Morais Neto (2001) houve avanços significativos nas técnicas de análise com o objetivo de produzir mapas cuja construção deve estar livre de artefatos relacionados à extensão da área geográfica e à população existente nas regiões estudadas. Usualmente, no entanto, o mapeamento de medidas de risco em epidemiologia é realizado através de dados agregados por área, com este tipo de análise perde-se a estrutura espacial detalhada dos dados, não conseguindo detectar variações em pequena escala (Shimakura et al, 2001).

Através da implementação de um Modelo Aditivo Generalizado que permite a inclusão de um fator de variação espacial no modelo (Wood, 2006), o que não é possível em outros tipos de modelos de Regressão mais usuais, pode-se entender o padrão de mortalidade infantil em uma região dentro de um contexto mais amplo, criando assim um processo pontual que consiste na modelagem de variáveis individuais e a localização pontual dos indivíduos identificando regiões de sobre-risco (Carvalho et al, 2004). Outro ponto forte que influência a escolha deste tipo de modelo é o fato de não haver necessidade de atendimento de suposições de normalidade, homocedasticidade e independência entre os erros, relacionamento que deve ocorrer em modelos paramétricos. Pode perceber que este tipo de modelo dentro de um processo espacial pontual tem uma grande aplicabilidade e potencial, já que claramente envolve vários campos interdisciplinares como a Estatística, Geografia, Sistemas de Informação, Saúde Pública entre outros, assim aprofunda o estudo e geração de informação e conhecimento na área de saúde publica dando uma base teórica e científica para que se façam planejamentos mais perspicazes provendo um método para descobrir padrões nos dados sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana ou ainda, numa análise não adequada que não venha suprir os interesses em questão.

O objetivo principal deste trabalho é identificar e mapear o padrão de distribuição espacial da mortalidade infantil em uma área urbana, e assim definir uma medida de risco para então investigar a significância da variação espacial e assim determinar e produzir mapas que identifiquem áreas de risco para dois componentes de mortalidade infantil: Mortalidade Neonatal e Mortalidade Pós-neonatal, controlando para potenciais fatores de risco individuais observáveis. Para isso será avaliada a variação espacial do risco através de um Modelo Aditivo Generalizado para identificar e controlar fatores individuais de risco e assim determinar se a variação espacial é constante ou não ao longo da região de estudo, ou seja, se a componente aditiva de variação no espaço é significativa ou não para explicar a mortalidade infantil nos dois casos citados acima.

2. REFERENCIAL TEÓRICO

O modelo de regressão usado neste estudo foi um Modelo Aditivo Generalizado (GAM), que nada mais é que uma extensão dos Modelos Lineares Generalizados (GLM) com um preditor linear envolvendo uma soma de funções não paramétricas ou funções suaves das covariáveis, em que o termo $X_i{}^t\beta = \sum X_{ij}\beta_j$ é substituído por $\sum f_j(X_{ij})$ com $f_j(X_{ij})$ sendo as funções suaves das covariáveis, com isso não é necessário assumir uma relação linear entre $g(\mu_i)$ e as covariáveis como no GLM, nem mesmo é necessário conhecer essa relação, mas é possível estimá-la a partir do conjunto de dados. O modelo GAM em geral tem uma estrutura como abaixo:

$$g(\mu_i) = X_i^t\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \qquad (2.1)$$

em que

$\mu_i = \mathrm{E}(Y_i)$ e Y_i segue uma distribuição da família Exponencial.

Sendo Y_i a variável resposta, X_i^t é o vetor transposto das covariáveis do modelo estritamente paramétrico, θ corresponde ao vetor dos parâmetros e f_j são as funções suaves das covariáveis.

Neste trabalho onde se têm dois tipos de eventos, recém-nascidos que morreram antes de completar um ano, denominados casos e os que sobreviveram a esta idade, controles. Sendo uma variável resposta do estudo do tipo binomial (assume 1 para caso e 0 para controle), dependente de diversas covariáveis tais como prematuridade, existência de doenças na gravidez, escolaridade da mãe, e incluindo sua localização no espaço, 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 (Druck et al, 2004). Será usada para isso uma abordagem semi-paramétrica também conhecida como abordagem de Modelos Aditivos Generalizados (GAM), pode-se com isso, controlando por fatores individuais de risco, a estimação espacial

do risco (Shimakura et al, 2001). O modelo semi-paramétrico será da seguinte forma:

$$\log\left\{\frac{p(s,x)}{1-p(s,x)}\right\} = \beta x_i + f(s_i), \qquad p(s,x) = p(Y=1 \mid s,x) \qquad (2.2)$$

em que:

- *Y* é a variável resposta, e tem a forma sim/não, zero/um (óbitos/não óbitos);
- a função de ligação da regressão é o *logito*, como usual para dados binomiais,
- x_i é o vetor de covariáveis;
- β é o vetor de parâmetros estimado pelo modelo, que no caso da regressão logística é a razão de chances (*odds ratio*) relacionada a cada covariável, ou seja, é o incremento de cada covariável ao logito;
- $f(s_i)$ é uma função suave desconhecida das coordenadas espaciais de s_i de casos e controles.

Dentre os métodos para regressão não paramétrica, os mais importantes são: método Kernel (Shimakura et al, 2001) e o via Splines (Wood, 2006). Neste estudo será usada a regressão via *Splines*.

2.1. REGRESSÃO VIA *SPLINES*

Splines é uma ferramenta proveniente do cálculo numérico, que vem ganhando atenção em estatística devido ao seu forte apelo adaptativo na aproximação de funções. Segundo Meneguette et al (2003) a curva definida por esse método pode ser descrita como sendo uma função por partes, cada qual um polinômio cúbico, de tal forma que ela e suas duas derivadas são sempre contínuas. A terceira derivada, entretanto, pode ter descontinuidade nos pontos x_i.

A representação da função suave pode ser mais bem introduzida considerando um modelo de regressão contendo uma função suave de uma covariável:

$$y_i = f(x_i) + \varepsilon_i, \qquad (2.3)$$

As funções *splines* estão associadas a uma partição do intervalo [a, b] onde se pretende trabalhar. A idéia está em dividir este intervalo em subintervalos menores, tal que:

$$I : a = x_0 < x_1 < ... < x_{m-1} < x_m = b$$

Em cada subintervalo $(x_{i-1}, x_i), i = 1,..., m$ as *splines* são polinômios de um determinado grau m. Esse procedimento produz um polinômio por partes s(x), que pode ser utilizado para aproximar a função procurada. Existe uma relação entre o grau dos "pedaços" dos polinômios e a ordem das derivadas exigidas nos pontos da partição. Assim, devem ser impostas algumas restrições na definição geral das *splines* para garantir a continuidade e suavidade de s(x), como a colocação dos nós (*knots*), os pontos de ligação entre os polinômios, dessas restrições e do acréscimo de peso a cada polinômio, surgem as bases de *splines* naturais e os *b-splines*.

Definição: uma função s(x) é chamada de *spline* de grau m, associada a uma partição do intervalo [a, b], se:

- s(x) é um polinômio de grau m em cada subintervalo $(x_{i-1}, x_i), i = 1,..., m$;
- s(x) tem m – 1 derivadas contínuas em cada x_i, e, portanto em [a, b].

Em que s(x) pode ser representada pela expressão:

s(x)= $\sum b_j(x)\beta_j$, para alguns valores desconhecidos dos parâmetros β_j e $b_j(x)$ é a j-ésima base da função que define o seu espaço. Por exemplo, suponha que s seja um polinômio de grau quatro, então o espaço dos polinômios de grau abaixo de quatro contém s. A base para este espaço é $b_1(x) = 1, b_2(x) = x, b_3(x) = x^2, b_4(x) = x^3, b_5(x) = x^4$. Então a expressão acima se tornará s(x)= $\beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5$.

Consideremos o modelo de regressão proposto em (2.3). Suponhamos que tenhamos uma função f(.) que estime a função g(x). Um critério de bondade de ajuste poderia ser dado por:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 \,. \tag{2.4}$$

Como estamos tratando com modelos numéricos inicialmente criados para interpolação, somente a bondade de ajuste não é um critério bom por si só, pois no caso de uma interpolação dos dados o ajuste teria uma bondade de ajuste excelente, mas seria muito pouco suave. Acrescentamos então um critério de suavidade, a dizer:

$$\int_a^b (f(x)^{(m)})^2 dx \,. \tag{2.5}$$

Juntando os critérios (2.4) e (2.5) em uma única equação, temos que o nosso problema se resume a minimizar:

$$A_\lambda(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int_a^b (f(x)^{(m)})^2 dx \tag{2.6}$$

para $\lambda > 0$, onde f(.) é gerada a partir de uma base de *splines.*

Notemos que λ determina o grau de suavidade da estimativa, controlando o quanto andamos na direção da interpolação dos dados ou na direção da suavização excessiva.

É bem conhecido que uma base de *splines* naturais cúbicos é a escolha mais apropriada para conseguir a minimização de (2.6). Uma escolha muito comum para as bases é o uso de *B-splines* cúbicos, pois são muito mais leves de se manipularem computacionalmente.

Fixada uma base de *splines*, o trabalho se resume em determinar o número e a localização dos pontos de junção (nós) dos polinômios cúbicos por partes.

Ficamos então com a tarefa da escolha do melhor parâmetro de suavização e da melhor combinação de nós. O método mais bem sucedido para realizar essa tarefa é o da Validação Cruzada Generalizada: a idéia consiste em tirar sucessivamente elementos da amostra e fazer uma estimativa do ponto retirado, obtendo-se um erro de predição. Procura-se então o conjunto de parâmetros que minimiza esse erro. Porém antes de nos aprofundarmos neste método, vamos

discorrer acerca das bases de *splines* que usaremos neste estudo, que são: *thin plate splines* e *tensor product.*

2.1.1. A *Thin plate splines*

Segundo Wood (2006) a *thin plate* é uma solução elegante e geral para o problema de estimar uma função suave de variáveis preditoras múltiplas, e segundo Wahba (2000) é uma generalização natural da *spline* polinomial univariada para duas ou mais dimensões via uma generalização do problema variacional em $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\int_a^b (f^{(m)}(x))^2 dx$. O problema variacional no espaço duplo Euclidiano é: encontrar *f* em um espaço $\boldsymbol{X}$ apropriado para minimizar

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i)))^2 + \lambda\sum_{\upsilon=0}^{m}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\binom{m}{\upsilon}\left(\frac{\partial^m f}{\partial x_1^{\upsilon}\partial x_2^{m-\upsilon}}\right)^2 dx_1 dx_2. \qquad (2.7)$$

Note que o limite na integral é $\pm\infty$. Se um limite finito é especificado, então um problema do valor de limite deve ser resolvido numericamente.

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

Uma grande característica da *thin plate splines* é a isotropia da penalidade das ondulações: ondulações em todas as direções são tratadas igualmente, com o ajuste inteiramente invariante para a rotação do sistema de coordenadas das variáveis preditoras.

2.1.2. O *Tensor product smooths*

Este método é usual no caso em que temos várias covariáveis em escalas diferentes, já que isto pode causar problemas na estimação. Então o *tensor product* associa penalidades separadas para cada covariável ao contrário do método *thin plat*e que utiliza uma penalidade simples como uma suavidade univariada, no entanto o *tensor product* (Wood, 2004) pode ser utilizado neste último caso também. O objetivo é fazer isto de uma maneira que permita a incorporação direta em modelos lineares mistos com a suavização separada por graus estimados para os diferentes sentidos das covariáveis.

2.2. A *SMOOTHING SPLINES*

Agora serão descrito os métodos para estimação do parâmetro suavizador.

2.2.1. Método da Validação Cruzada (CV)

O método da validação cruzada propõe um procedimento automatizado para estimação do parâmetro λ. Seja $f_\lambda^{[k]}$ a função que minimiza:

$$\frac{1}{n}\sum_{i\neq k}(y_i - f(x_i))^2 + \lambda\int_a^b (f^{(n)})^2 \qquad (2.8)$$

Ou seja, procuramos uma otimização retirando cada um dos pontos k.

A função de validação cruzada ordinária V0(λ) é definida por:

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda^{[k]}(x_k))^2 \qquad (2.9)$$

Utilizando então algumas identidades obtêm-se uma forma simplificada de V0, a qual tem um custo computacional muito mais baixo, a dizer:

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f(x_k))^2/(1-h_k(\lambda))^2 \qquad (2.10)$$

onde $h_k(\lambda)$ é o k-ésimo elemento da matriz H_λ, onde $f_\lambda = H_\lambda y$.

2.2.2. Método da Validação Cruzada Generalizada (GCV)

O método da validação cruzada generalizada se baseia no método da validação cruzada, mas tem algumas vantagens. Ele é mais barato computacionalmente e possui algumas propriedades teóricas que não são possíveis de se obterem com o método anterior. A função de validação cruzada generalizada segue abaixo:

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda(x_k))^2/(1-\overline{h}_k(\lambda))^2 = \frac{\frac{1}{n}\|(I-H_\lambda)y\|^2}{\left(\frac{1}{n}tr(I-H_\lambda)^2\right)} \text{ onde } \overline{h}(\lambda) = (1/n)tr(H_\lambda) \quad (2.11)$$

É importante ainda notar que os métodos de validação cruzada estão bem definidos para um conjunto de pontos $\{x_i\}_{i=1}^n$ distintos, e que portanto deve-se tomar cuidado na sua implementação para eliminar grupos de pontos não distintos antes de realizar o procedimento de otimização.

2.2.3. Método UBRE

O estimador UBRE (*Un-Biased Risk Estimator*) (Wood,2006), é usado nos casos em que a escala do parâmetro é supostamente conhecida, para minimizar o erro quadrático médio ligado á estimação por Cp de *Mallow* (Montgomery). Se comparado o método UBRE com o GCV, que é usado no caso em que a escala do parâmetro é supostamente desconhecida, com dados simulados, constata-se que os escores de ambos tendem a estarem completamente próximos, e estão em um acordo mais próximo um com o outro do que com o erro quadrático médio observado mais a variância. Outro detalhe interessante sobre este estimador, que efetivamente trata-se exatamente de uma transformação linear do Critério de *Akaike* (AIC), (Pérez, 2004).

2.3. ÁRVORE DE CLASSIFICAÇÃO

Segundo Braga L. P. V. (2005), a Árvore de Classificação é um procedimento hierárquico para predizer a classe de um objeto com base em suas variáveis preditoras e também para definir classes. Uma das principais características deste método é o seu tipo de reapresentação: uma estrutura hierárquica que traduz uma árvore invertida que se desenvolve da raiz para as folhas de modo que cada nível da árvore desenvolve um problema complexo de classificação em subproblemas mais simples, assim podendo ser feitas previsões com menos riscos.

Uma das formas de obtermos uma Árvore de Classificação é considerarmos um conjunto de dados D, em que, $D=\{x_1, x_2, x_{3,...}, x_n\}$ e considerarmos igualmente as classes definidas por C, em que, C está definida da seguinte forma, $C=\{c_1, c_2, c_{3,...}, c_m\}$, com $m<n$. O objetivo será estabelecer uma ralação segundo uma função f definida como $f : D \rightarrow C$, em que, cada vetor x de D, cujas coordenadas são valores assumidos pelas variáveis explicativas que descrevem a amostra para o caso x, a uma classe de C. Podemos definir então a regra de classificação ou classificador como uma função $f(x)$ definida em D, que para todo o x pertence a D, $f(x)$ pertence a uma das classes $c_1, c_2, c_{3,...}, c_m$, ou equivalentemente, uma partição de D em m subconjuntos disjuntos $A_1, A_2, A_{3,....}, A_k$, com:

$$D= U\, A_j \qquad (2.12)$$

tal que, para todo x pertencente a A_j a classe prevista é a j.

As Árvores de Classificação podem ser usadas com objetivos diferentes, de acordo com o problema que se pretende resolver. Neste trabalho a utilização deste método, além de identificar padrão da mortalidade infantil dentro das componentes neonatal e pós-neonatal, foi utilizado para identificar interações entre as variáveis independentes e assim testar estas interações no modelo GAM e também recategorizar, de forma binária, as cováriáveis que apresentavam mais de duas categorias. Há vários algoritmos na literatura para construção de modelos de Árvores de Classificação. O método adotado por esses algoritmos consiste na divisão recursiva do conjunto de observações em subgrupos filhos. Neste trabalho foi aplicado o algoritmo *CHAID* (Rodrigues, 2005).

3. APLICAÇÃO: MORTALIDADE INFANTIL EM CURITIBA – PR

Neste capítulo serão descritos os dados utilizados e os resultados obtidos na análise da distribuição espacial e de fatores de risco dos óbitos de crianças com menos de 28 dias (Mortalidade Neonatal) e dos óbitos de crianças de 28 dias a 12 meses incompletos (Mortalidade Pós-neonatal) dentro do espaço geográfico urbano de Curitiba em 2004.

2.4. APRESENTAÇÃO DOS DADOS

Para a realização deste trabalho foram utilizados os bancos de dados: SIM (Sistema Sobre Mortalidade) e SINASC (Sistema de Informação Sobre Nascidos Vivos), que por sua vez foram cedidos pela Secretaria Municipal de Saúde do Município de Curitiba, sendo ambos referentes ao ano de 2004.

O relacionamento dos dois bancos de dados foi realizado através de um software específico de manipulação de bancos dados de nome *ACL*. A linkagem foi feita pelo nome da mãe dos casos. Da mesma, o georreferenciamento de nascidos e óbitos foi realizado pelo SIG implantado no Instituto de Pesquisa e Planejamento Urbano de Curitiba – IPPUC.

Antes de iniciar as análises foram selecionados todos os óbitos que ocorreram com até 12 meses incompletos de vida em 2004 e que haviam nascidos neste mesmo ano, que foram denominados casos, informações essas que foram extraídas do SIM. O próximo passo foi relacionar estes casos com o SINASC.

Para controle da heterogeneidade da distribuição espacial da população, foram selecionados aleatoriamente controles no Banco do SINASC. Foi adotado um esquema de amostragem de 2:1, ou seja, dois nascidos para cada óbito.

De um total de 239 óbitos com até um ano de vida registrados em 2004 pelo sistema de nascidos vivos, foram retirados 54 por falta de informação e pelo fato de o local de residência não ser dentro do município de Curitiba ou impossibilidade de georreferenciamento, sendo selecionados, 136 registros Neonatal e 49 registros Pós-neonatal.

As informações ou variáveis analisadas foram as seguintes:

- Y, variável de interesse estudada: 1 (óbito) e 0 (caso contrário) .

- XCOORD, YCOORD, localização da residência, que se apresenta em coordenada espacial;
- PESO, peso da criança ao nascer em gramas;
- GESTACAO, duração da gestação em semanas, codificada da seguinte forma: 1 (até 36 semanas), 0 (acima de 37 semanas);
- SEXO, sexo da criança: M (Masculino) e F (Feminino);
- PARTO tipo de parto: 1 (Normal) e 2 (Cesário);
- IDADEMAE, idade da mãe em anos completos;
- ESCMAE, grau de instrução da mãe: 1 (1º grau incompleto), 0 (1º grau completo);
- QTDFILMORT, quantidade de filhos mortos;
- QTDFILVIVO, quantidade de filhos vivos;

Para o cálculo do risco serão utilizados controles selecionados aleatoriamente no banco de dados do SINASC, entre crianças que nasceram no mesmo ano de ocorrência dos óbitos, sendo considerados representantes da distribuição espacial da população de risco.

O modelo estatístico foi ajustado através do pacote estatístico R, de código aberto, com a utilização da biblioteca *mgcv*. As árvores de classificação, que também foram aqui utilizadas, são provenientes de um pacote específico de mineração de dados chamado Knowledge Studio versão 5.2 (Angoss, 2006).

2.5. RESULTADOS

Este tópico será dividido entre apresentação dos resultados da análise descritiva que foi feita com o intuito de descrever, sumarizar e melhor compreender seu comportamento e identificação de padrões implícitos, no conjunto estudado, e análise preditiva que, através de um Modelo Aditivo Generalizado, teve como foco a identificação do padrão da distribuição espacial da mortalidade infantil, divididos em componentes de mortalidade pós-neonatal e neonatal.

2.5.1. Resultados da análise descritiva

Primeiramente analisaram-se as covariáveis. Para as variáveis, peso da criança ao nascer e idade materna no nascimento da criança, que são contínuas, foi criado um sumário e histogramas de freqüência.

TABELA 3.1 - RESUMO NUMÉRICO DA VARIÁVEL PESO EM GRAMAS SEGUNDO CASOS E CONTROLES – 2004

RESUMO NUMÉRICO	PESO (g)	
	Caso	Controle
Neonatal		
Média	1.777	3.138
Intervalo de Confiança para média[(1)]	(1.597,02 ; 1.956,28)	(3.078,19 ; 3.197,92)
Desvio Padrão	1.059,21	502,43
Mínimo	380	1.410
1º quartil	817,5	2.875
Mediana	1.485	3.180
3º quartil	2.679	3.445
Máximo	4.220	4.325
Pós-neonatal		
Média	2.566	3.142
Intervalo de Confiança para média[(1)]	(2.316,03 ; 2.816,81)	(3.038,43 ; 3.244,83)
Desvio Padrão	871,73	514,76
Mínimo	690	970
1º quartil	2.110	2.886
Mediana	2.690	3.148
3º quartil	3.165	3.390
Máximo	4.455	4.355

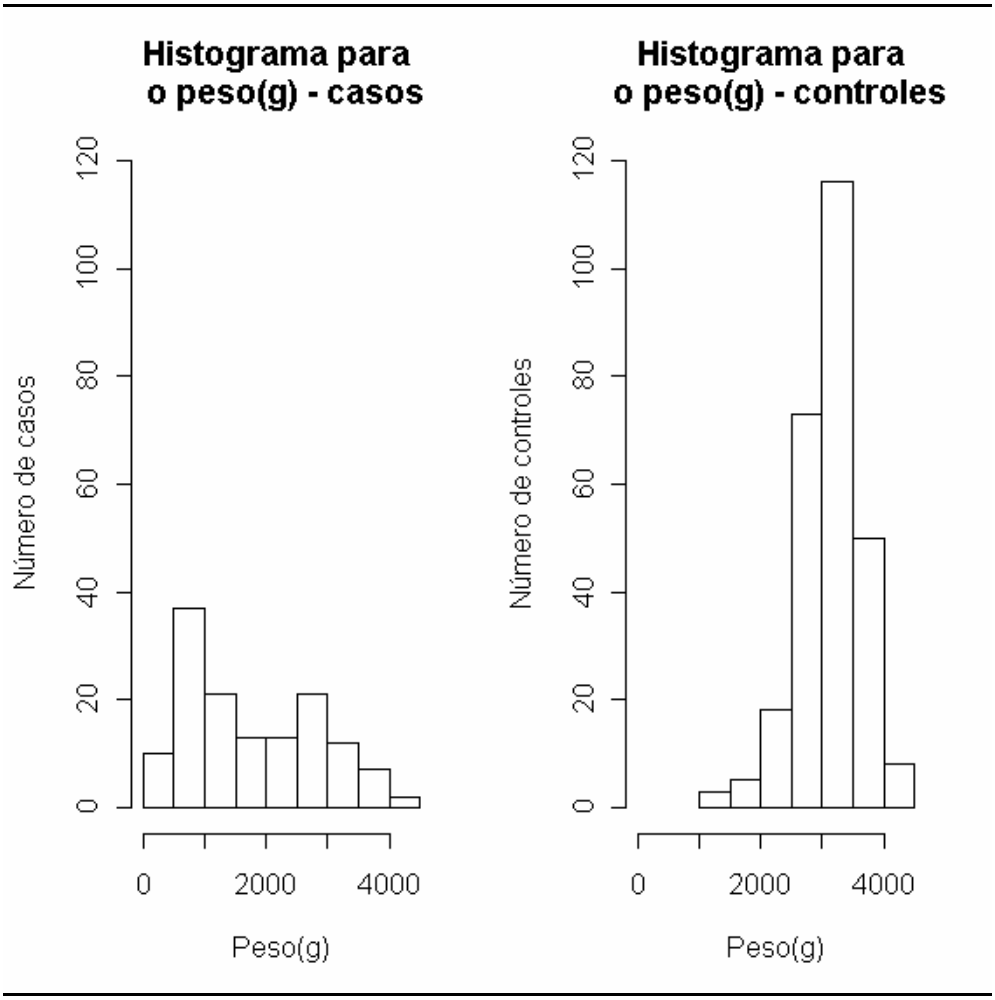
FONTES: SIM, SINASC
(1) Intervalo de Confiança ao nível de 95% de significância.

Constatou-se na Tabela 3.1 que o peso médio das crianças que foram a óbito até 28 dias de vida (neonatal) é de 1.777 g com um desvio padrão alto,de 1.059,21 g indicando uma grande dispersão e o peso médio das crianças que entraram em óbito entre 28 dias e 12 meses (pós-neonatal) foi maior, 2.566 g com desvio padrão 871,5 g. Observou-se também que a maior diferença entre o peso médio das crianças casos e controles deu-se no grupo neonatal, em que os controles tiveram um peso médio de 3.138 g com um desvio padrão 502,43. Viu-se que existe uma

maior dispersão nos dados nos casos dentro do grupo neonatal. Os outros grupos não se mostraram tão dispersos.

Na Figura 3.1 percebe-se que apenas os controles se distribuem quase que simetricamente em torno do peso médio. Os casos, além da assimetria, mostram uma dispersão maior entre os valores dos pesos, confirmando assim o que foi visto no resumo numérico na Tabela 3.1.

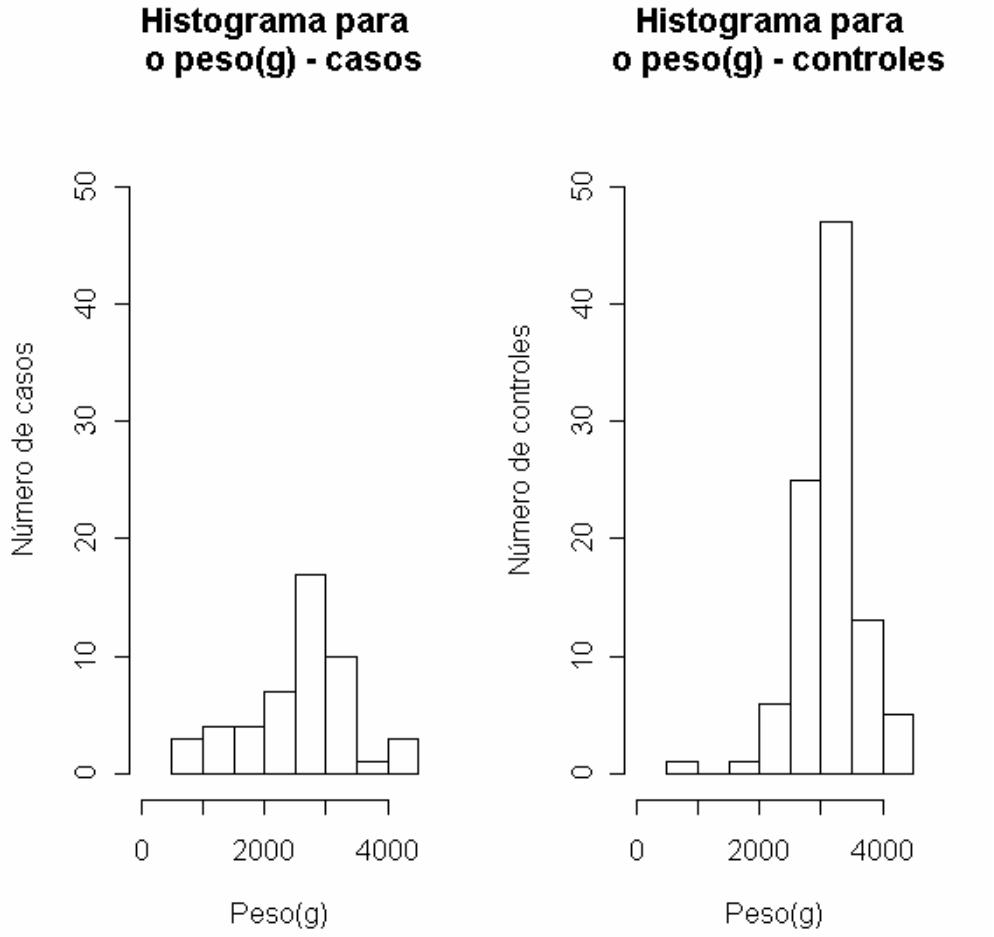
FIGURA 3.1 – HISTOGRAMAS DE FREQUÊNCIA PARA A VARIÁVEL PESO SEGUNDO CASOS E CONTROLES NEONATAL



FONTES: SIM, SINASC

Na Figura 3.2 constata-se que no pós-neonatal, tanto para os casos quanto para os controles, o peso distribui-se quase que simetricamente em torno do peso médio.

FIGURA 3.2 – HISTOGRAMAS DE FREQUÊNCIA PARA A VARIÁVEL PESO SEGUNDO CASOS E CONTROLES PÓS-NEONATAL



FONTES: SIM, SINASC

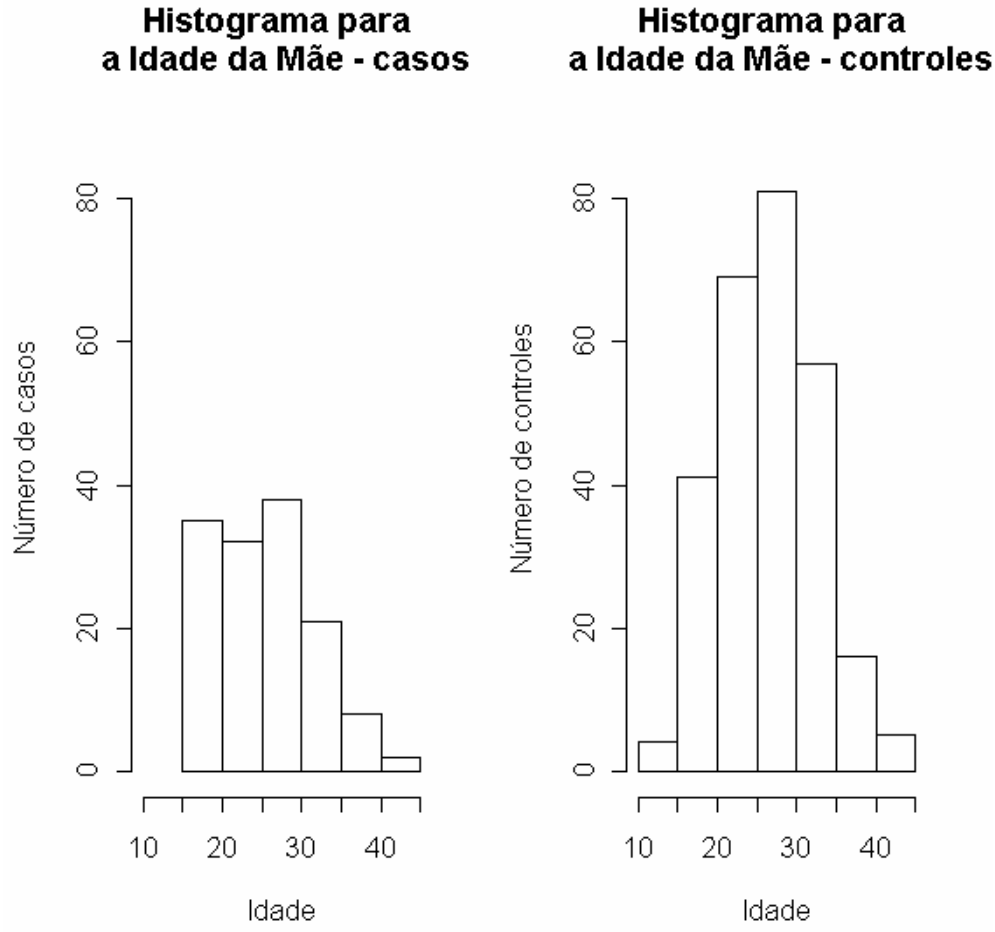
Na Tabela 3.2, nota-se que as diferenças entre as idades médias das mães, tanto para os casos quanto para os controles, foi de aproximadamente de 1 ano a mais para os controles, este comportamento foi análogo aos dois componentes estudados para a mortalidade. Em geral, percebe-se que cerca de 75% das mães têm idade entre 13 a 31 anos.

TABELA 3.2 - RESUMO NUMÉRICO DA VARIÁVEL IDADE DA MÃE SEGUNDO CASOS E CONTROLES – 2004

RESUMO NUMÉRICO	IDADE DA MÃE	
	Caso	Controle
Neonatal		
Média	25,74	27,03
Intervalo de Confiança para média[(1)]	(24,65 ; 26,82)	(26,3 ; 27.75)
Desvio Padrão	6,4	6,11
Mínimo	15	13
1º quartil	20	22
Mediana	26	27
3º quartil	30	31
Máximo	41	44
Pós-neonatal		
Média	25	26,12
Intervalo de Confiança para média[(1)]	(22,79 ; 27,17)	(24,92 ; 27,32)
Desvio Padrão	7,63	5,97
Mínimo	14	14
1º quartil	19	22
Mediana	23	26,5
3º quartil	30	30
Máximo	44	41

FONTES: SIM, SINASC
(1) Intervalo de Confiança ao nível de 95% de significância.

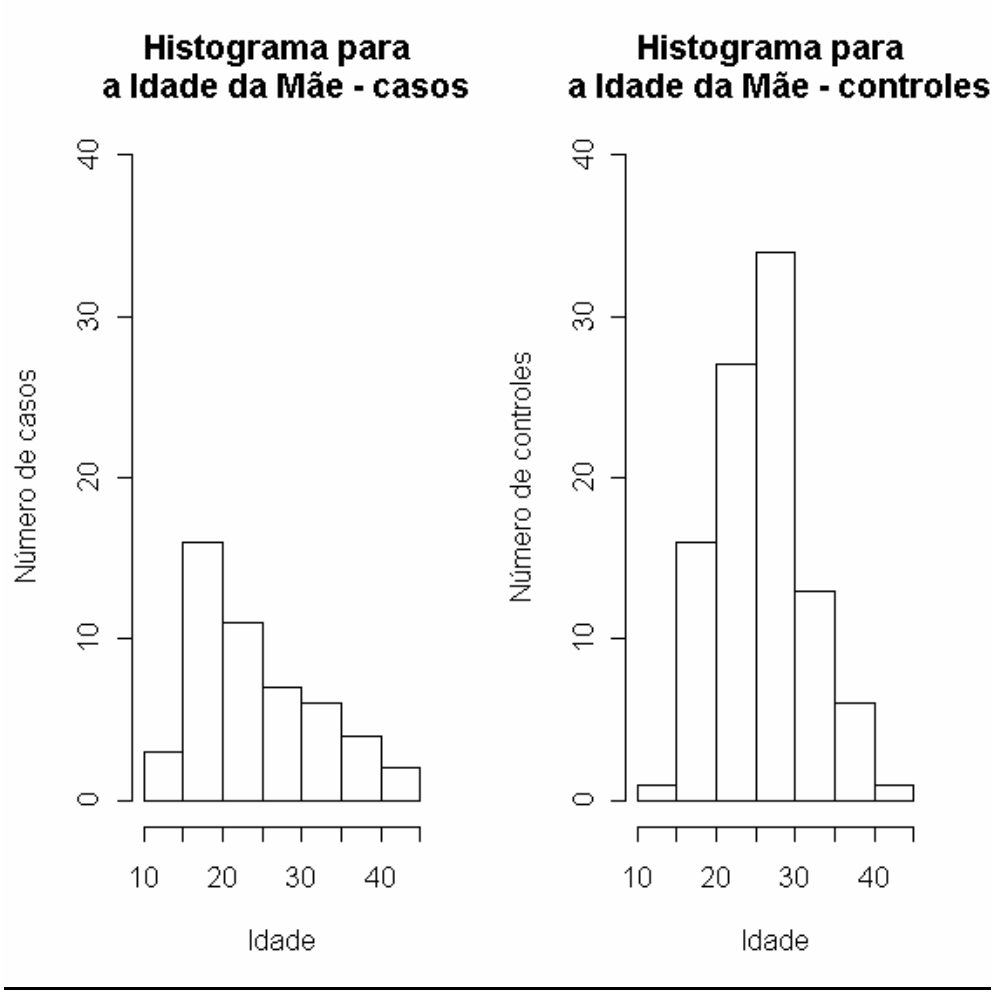
FIGURA 3.3 – HISTOGRAMAS DE FREQUÊNCIA PARA A VARIÁVEL IDADE DA MÃE SEGUNDO CASOS E CONTROLES NEONATAL



FONTES: SIM, SINASC

Na Figura 3.3, nota-se que apenas as mães para o grupo controle têm uma distribuição quase simétrica em torno da idade média, o que não ocorre nos casos em que a concentração se apresenta em uma classe de mães mais jovens, para componente neonatal.

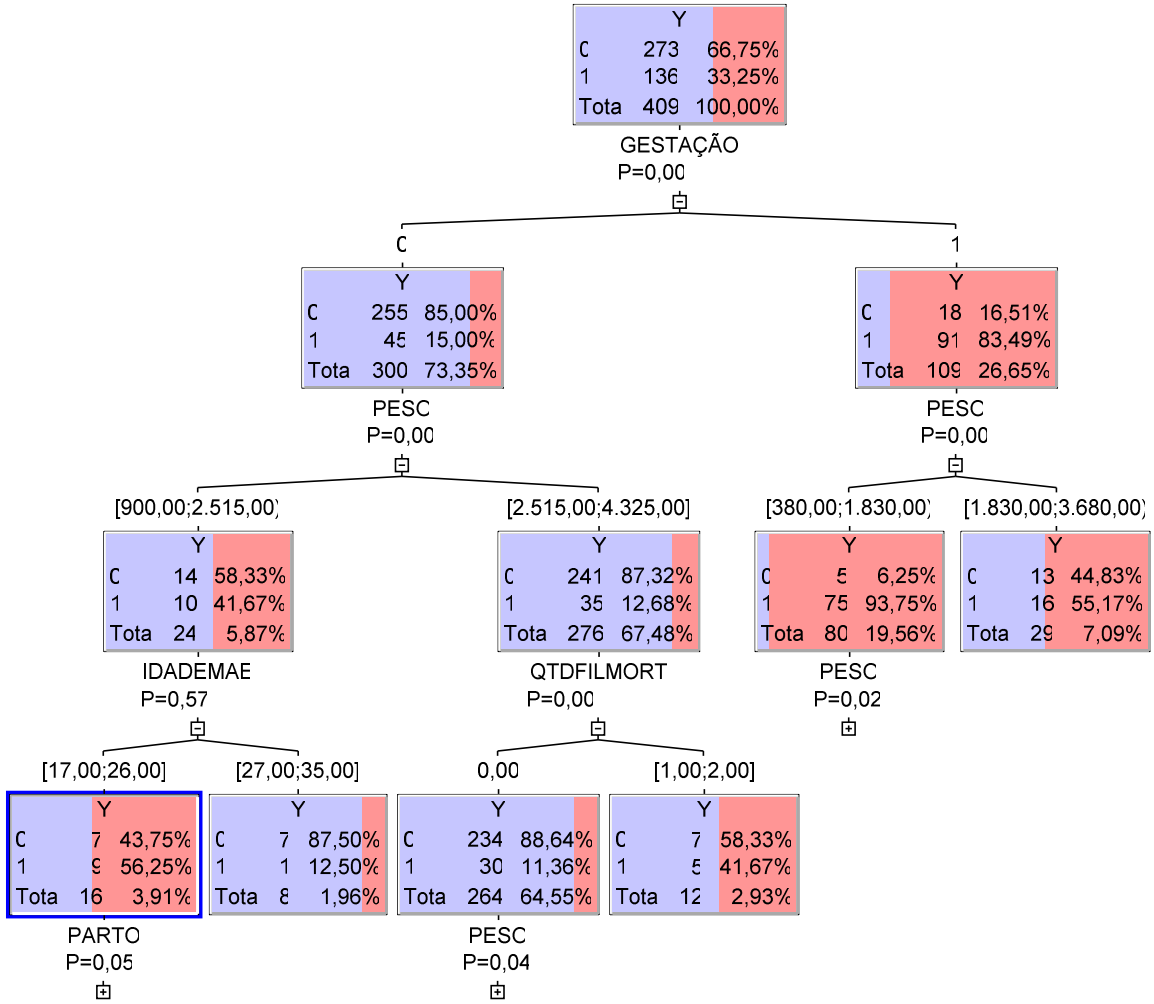
FIGURA 3.4 – HISTOGRAMAS DE FREQUÊNCIA PARA A VARIÁVEL IDADE DA MÃE SEGUNDO CASOS E CONTROLES PÓS-NEONATAL



FONTES: SIM, SINASC

A idade das mães na componente Pós-neonatal representada na Figura 3.4 apresenta uma distribuição aparentemente simétrica em torno da idade média das mães para os controles, o que não se observa nos casos, já que sua distribuição, aparenta um formato similar da Distribuição Gama, ou seja, há muitas mães jovens.

FIGURA 3.5 – ÁRVORE DE CLASSIFICAÇÃO PARA MORTALIDADE NEONATAL



Na Figura 3.5 observa-se que a covariável duração da gestação é a mais importante para explicar a mortalidade infantil para a componente neonatal interagindo com as covariáveis, peso, idade da mãe e quantidade de filho morto.

A cor rosa representa a proporção de óbitos e a cor azul corresponde aos controles, pode-se então constatar que crianças que tiveram uma gestação até 36 semanas e com peso até 1.830 gramas têm mais propensão a entrarem em óbito neonatal. Para a componente pós-neonatal (Figura 3.6) as cores têm as mesmas atribuições que foram citadas anteriormente. A covariável gestação também se apresentou a mais importante para explicar o óbito em Curitiba em 2004, interagindo com as variáveis, quantidade de filhos vivos, idade da mãe e peso. Percebe-se, assim como para a componente neonatal que crianças que tiveram uma gestação até 36 semanas têm mais propensão a morrer no pós-neonatal.

FIGURA 3.6 – ÁRVORE DE CLASSIFICAÇÃO PARA MORTALIDADE PÓS-NEONATAL

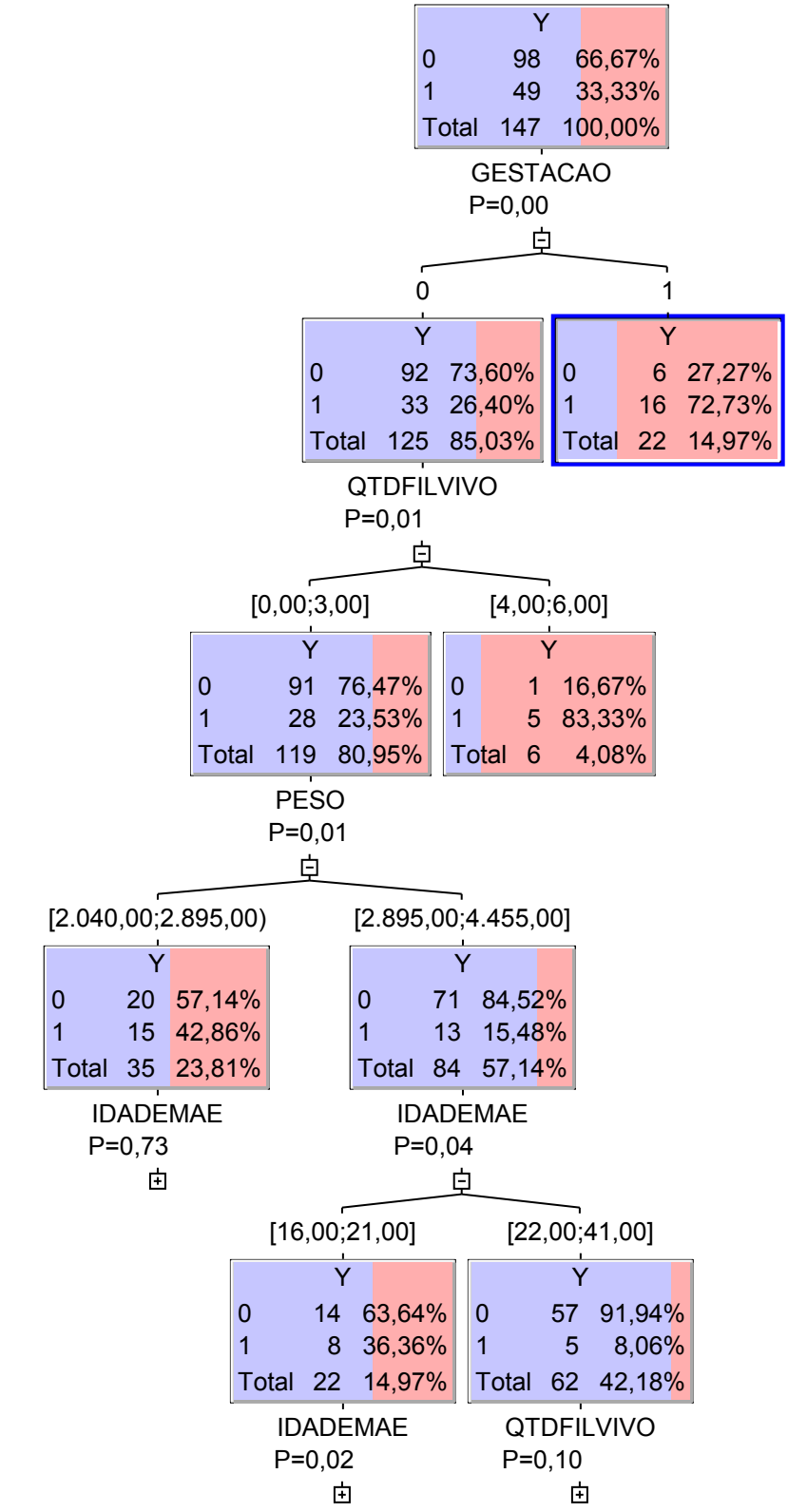
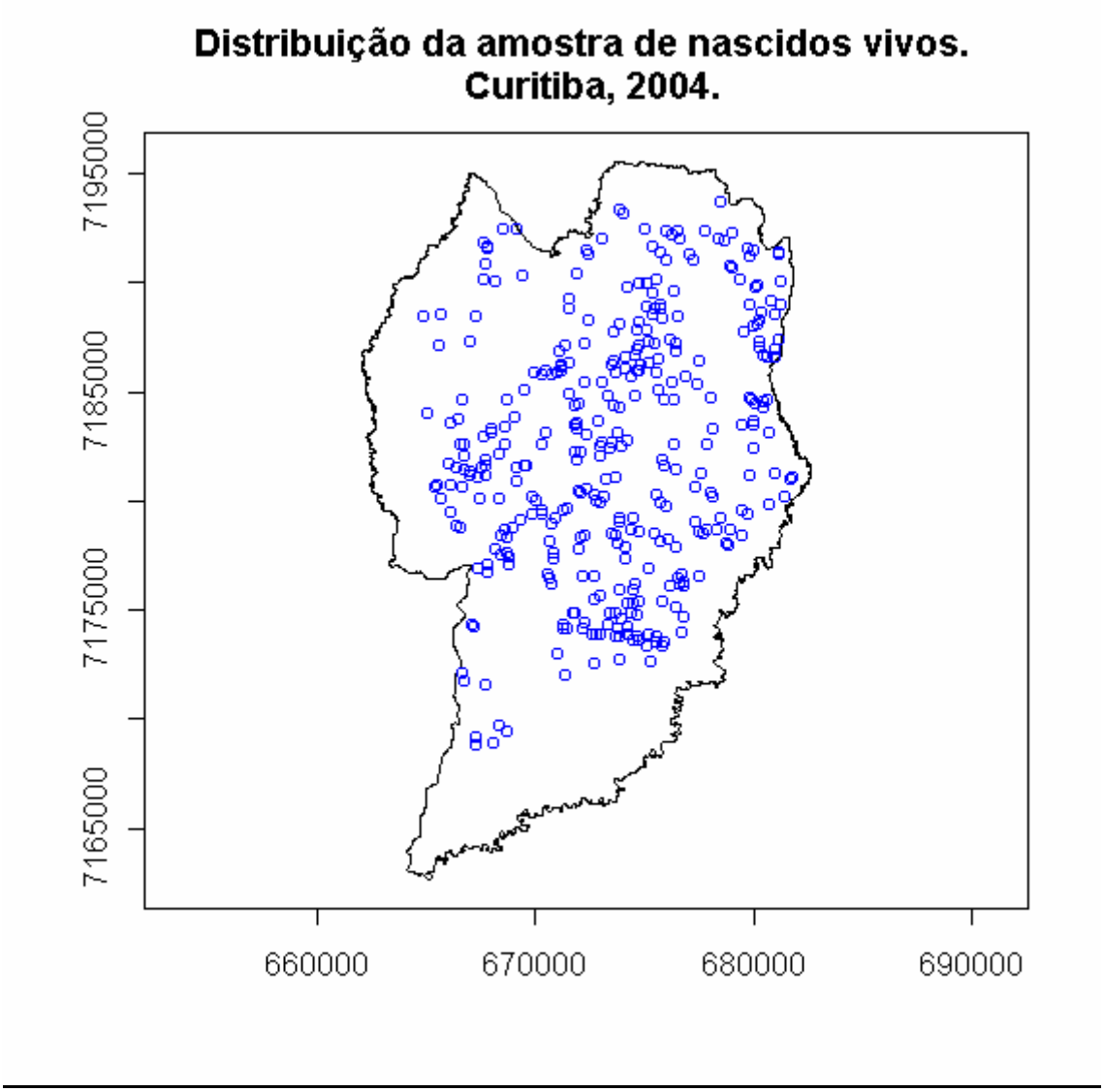


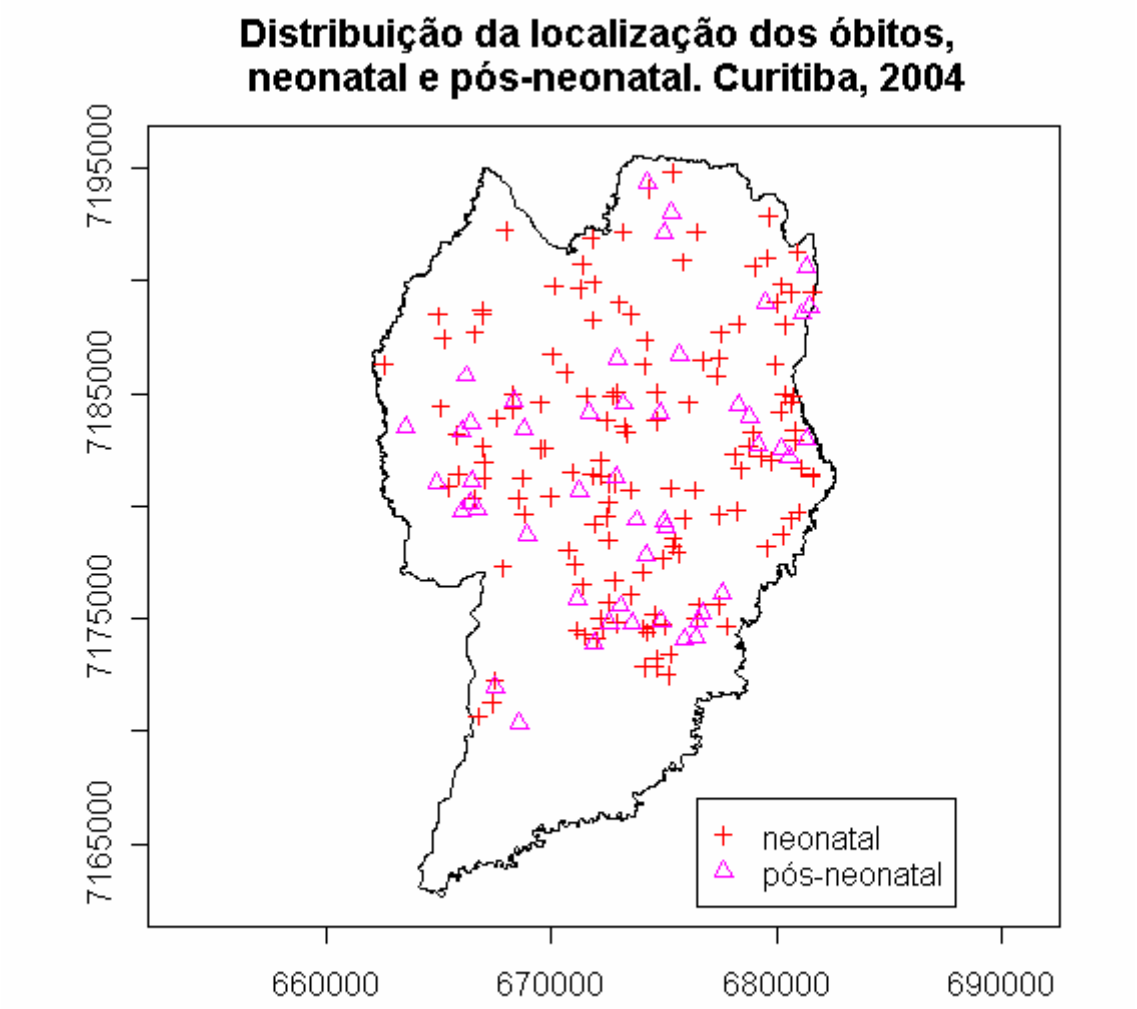
FIGURA 3.7 – MAPA DE CURITIBA COM OS CONTROLES



FONTES: SIM, SINASC, IPPUC.

Nas Figuras 3.7 e 3.8 são apresentados respectivamente: a localização pontual da residência da amostra de nascidos vivos registrados no SINASC e a localização dos óbitos infantis (neonatal e pós-neonatal) em Curitiba em 2004, com isso nota-se a heterogeneidade da distribuição da população em risco (Figura 3.7) e conseqüentemente a heterogeneidade da os casos também (Figura 3.8).

FIGURA 3.8 – MAPA DE CURITIBA COM OS CASOS



FONTES: SIM, SINASC, IPPUC.

2.5.2. Resultados da análise preditiva

Dividiu-se, como dito anteriormente, em mortalidade neonatal e pós-neonatal, para a análise da variação espacial para o risco da mortalidade infantil em Curitiba no ano de 2004.

Através do método de seleção *Stepwise* (Giolo, 2003), para a identificação de variáveis explicativas significativas para o risco de morte, foram ajustados diversos modelos GAM e dentre eles escolhido o melhor para cada uma das componentes. Para o ajuste dos modelos GAM e estimação da função suave do termo não-

paramétrico foram definidas um número de nós para a base igual a 10, número este que é arbitrário.

Para a mortalidade neonatal o modelo final adotado teve como fatores de risco significativos: peso, quantidade de filhos vivos, idade da mãe, gestação e parto.

Segue abaixo o modelo:

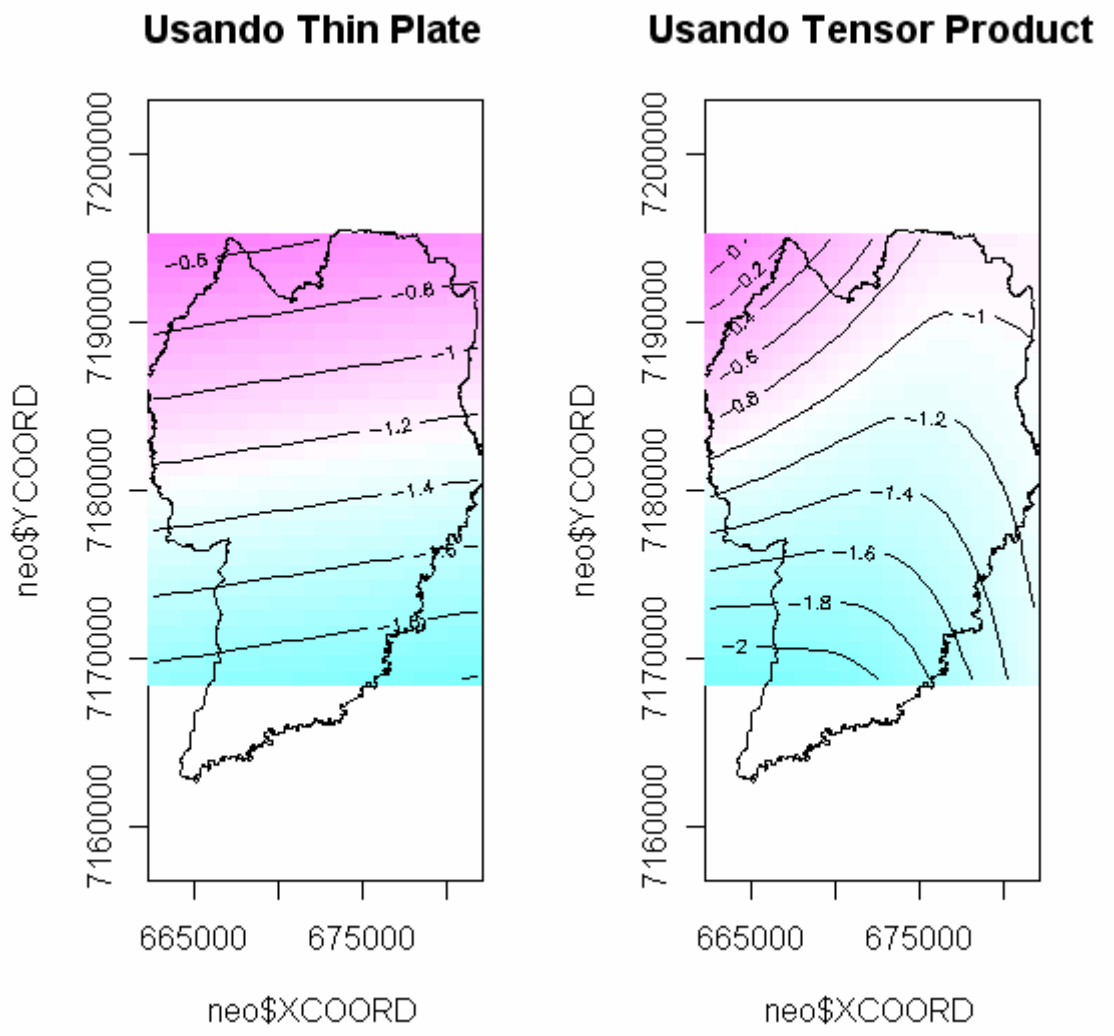
$$\log\left\{\frac{p(s,x)}{1-p(s,x)}\right\} = \beta_0 + \beta_1 * peso + \beta_2 * gestação + \beta_3 * \text{idade mãe} + \beta_4 * qtdfilvivo + \beta_5 * parto + f(s_i)$$

TABELA 3.3 – ESTIMATIVAS DOS EFEITOS DAS COVARIÁVEIS USANDO AS BASES *THIN PLATE* E *TENSOR PRODUCT* PARA NEONATAL

BASE/ FATOR	Estimativa	Erro padrão	P-valor
Thin plate			
Peso	-0,001	0,000	0,000
Gestação	1,405	0,494	0,004
Idade da mãe	-0,085	0,028	0,002
Quant. Filhos vivos	0,384	0,136	0,005
Parto	0,787	0,327	0,016
Tensor product			
Peso	-0,001	0,000	0,000
Gestação	1,424	0,495	0,004
Idade da mãe	-0,085	0,028	0,002
Quant. Filhos vivos	0,389	0,136	0,004
Parto	0,789	0,327	0,016

Na Tabela 3.3 são apresentadas às estimativas dos parâmetros do modelo usando *thin plate* e *tensor product* controlando-se pelo fator espacial, nota-se que não há diferença expressiva entre as bases. Dentre as covariáveis confirma-se a influência do baixo peso ao nascer na determinação do óbito, assim como a baixa idade da mãe. A covariável gestação, em relação ao nascimento antes de 37 semanas, apresentou-se como um fator de risco, o parto normal se mostrou como um fator de proteção e quanto maior o número de filhos a mãe possui maior o risco para o óbito.

FIGURA 3.9 – MAPAS DE RISCO PARA MORTALIDADE NEONATAL EM CURITIBA, PARANÁ, 2004.



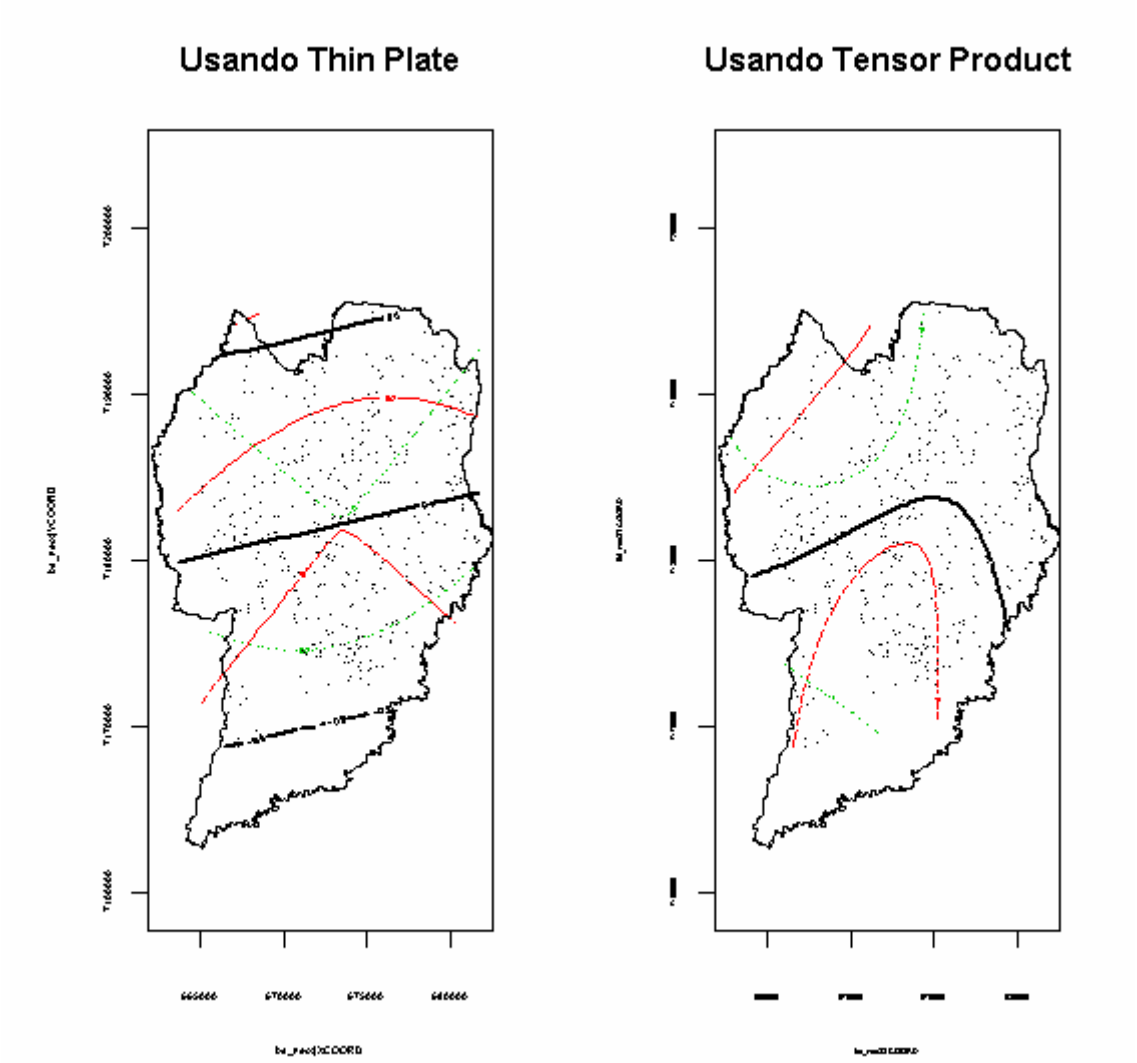
FONTES: SIM, SINASC, IPPUC.

A estimativa do termo não-paramétrico *f(s)*, que é interpretada como surpefície de risco estimada controlando-se por fatores individuais são representados nas Figuras 3.9 e 3.10, para a componente neonatal.

A análise não apresentou variação espacial significativa no risco de mortalidade neonatal sobre a cidade de Curitiba em 2004, tanto usando a *Thin Plate* (p-valor=0,399) quanto usando o *Tensor Product* (p-valor=0,571), ou seja, o risco de óbito não aparenta variar significativamente ao longo do espaço para a componente neonatal embora ao Norte pareça ter um risco mais elevado. A Figura 3.10 mostra a suavidade estimada segundo seus graus de liberdades no espaço, sendo os

pontinhos os resíduos parciais, a linha verde a suavidade estimada mais o erro padrão e a vermelha a suavidade menos o erro padrão.

FIGURA 3.10 – MAPAS DE SUAVIDADE ESTIMADA PARA MORTALIDADE NEONATAL EM CURITIBA, PARANÁ, 2004



FONTES: SIM, SINASC, IPPUC.

Para a mortalidade pós-neonatal houve diferença entre os ajustes quando comparadas as bases, isto é, não ocorreu convergência entre os modelos como observado na componente neonatal.

Segue abaixo os modelos ajustados:

Thin Plate

$$\log\left\{\frac{p(s,x)}{1-p(s,x)}\right\} = \beta_0 + \beta_1 * peso + \beta_2 * idade\ mãe + \beta_3 * qtdfilvivo + f(s_i)$$

Tensor Product

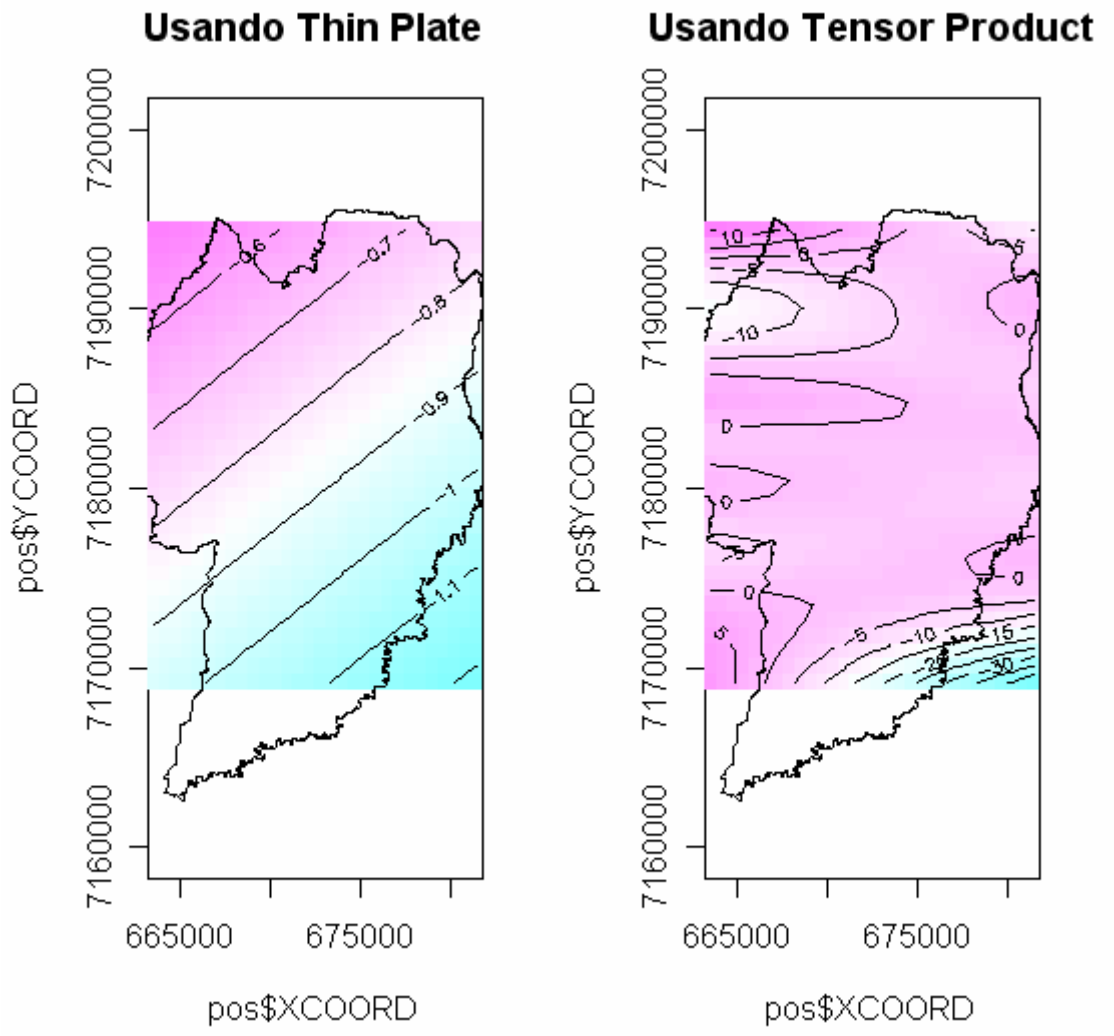
$$\log\left\{\frac{p(s,x)}{1-p(s,x)}\right\} = \beta_0 + \beta_1 * peso + \beta_2 * idade\ mãe + \beta_3 * qtdfilvivo + \beta_4 * idade\ mãe * qtdfilvivo + f(s_i)$$

TABELA 3.4 – ESTIMATIVAS DOS EFEITOS DAS COVARIÁVEIS USANDO AS BASES THIN PLATE E TENSOR PRODUCT PARA PÓS-NEONATAL

BASE / FATOR	Estimativa	Erro padrão	P-valor
Thin plate			
Peso	-0,001	0,000	0,000
Idade da mãe	-0,095	0,039	0,015
Quant. Filhos vivos	0,706	0,205	0,001
Tensor product			
Peso	-0,002	0,000	0,000
Idade da mãe	-0,242	0,072	0,001
Quant. Filhos vivos	-1.522	1,146	0,184
Idade da mãe: Quant. Filhos vivos	0,086	0,041	0,035

Assim como foi visto para a componente neonatal as covariáveis peso e idade da mãe têm influência sobre o óbito (Tabela 3.4). Ao contrário da *Thin Plate* a interação entre idade da mãe e quantidade de filhos vivos mostrou-se significativa, ou seja, quanto maior a idade da mãe e quantidade de filhos vivos maior a propensão ao óbito.

FIGURA 3.11 – MAPAS DE RISCO PARA MORTALIDADE PÓS-NEONATAL EM CURITIBA, PARANÁ, 2004

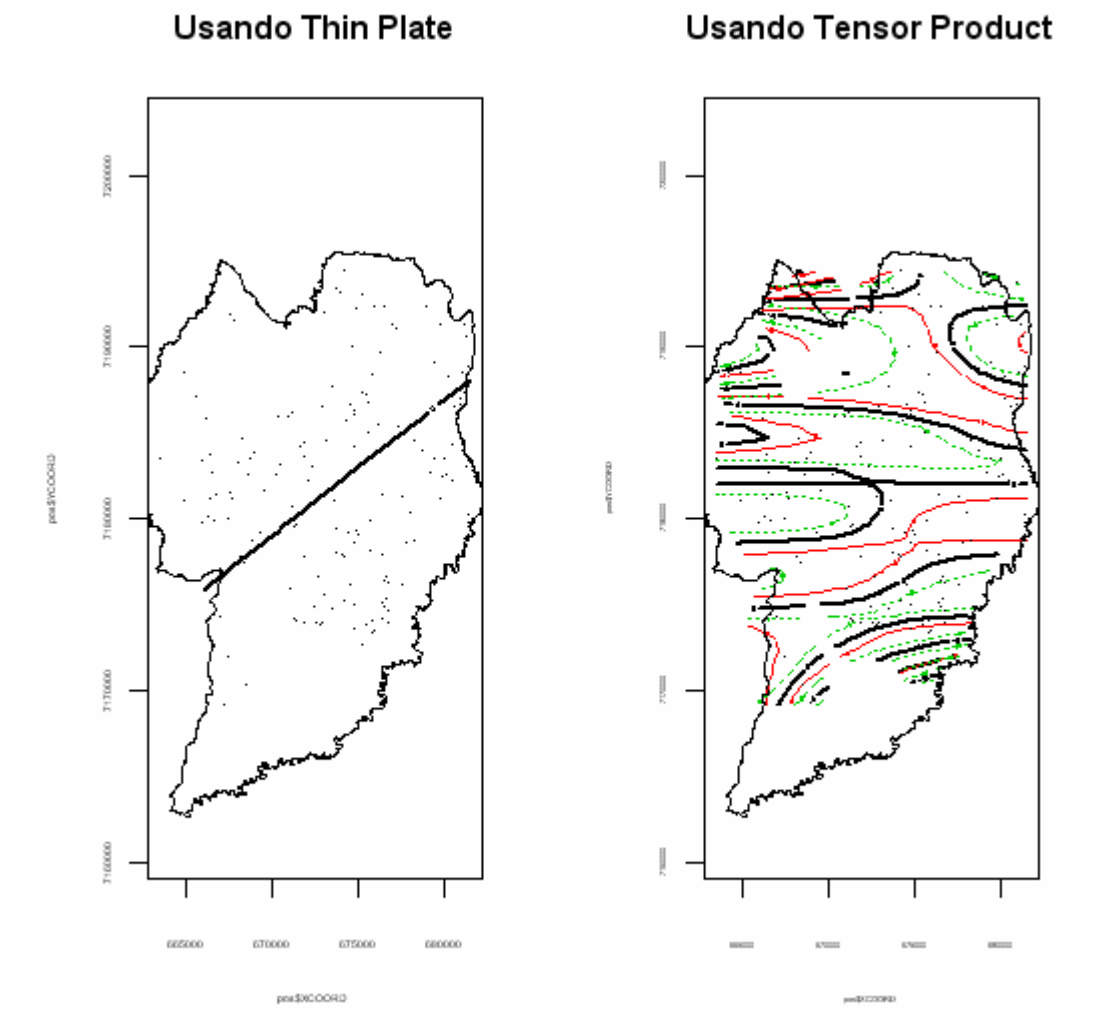


FONTES: SIM, SINASC, IPPUC.

A estimativa do termo não-paramétrico *f(s)*, que é interpretada como surpefície de risco estimada controlando-se por fatores individuais são representados nas Figuras 3.11 e 3.12, para a componente pós-neonatal.

Para este caso a análise apresentou uma variação espacial significativa quando usado o *Tensor Product* (p-valor=0,0376), o que não ocorreu quando usado a *Thin Plate* (p-valor=0,944). Confirma-se a capacidade da bas e *Tensor Product* detectar com mais sensibilidade a variação do risco quando esta é mais tênue ao longo do espaço. Verificando duas áreas que podem ser de proteção, uma ao extremo Noroeste e outra ao extremo Sudeste da cidade de Curitiba em 2004.

FIGURA 3.12 – MAPAS DE SUAVIDADE ESTIMADA PARA MORTALIDADE PÓS-NEONATAL EM CURITIBA, PARANÁ, 2004



FONTES: SIM, SINASC, IPPUC.

4. CONSIDERAÇÕES FINAIS

As análises mostraram a grande influência do baixo peso da criança ao nascer para ambas componentes de mortalidade infantil que aqui foram estudadas, assim como a baixa idade da mãe. O que foi observado neste estudo, diz respeito à baixa escolaridade da mãe como não sendo um fator de risco, ao contrário de estudos de mesma natureza que apresentou esta variável sendo significativa para a determinação de óbitos (Shimakura et al, 2001). Outro fator que se deve ressaltar como fator de risco é a quantidade de filhos vivos, isto é, quanto mais filhos a mãe possui mais risco da criança entrar em óbito, possivelmente pela qualidade da atenção dedicada ás crianças. Uma diferença entre o risco de óbito entre as duas componentes aqui estudadas se mostrou em relação a dois fatores que são idade gestacional e tipo de parto e a interação entre a idade da mãe e quantidade de filhos vivos, ratificando assim que a mortalidade neonatal está mais associada a fatores de ordem congênita e complicações no parto, tanto que o tipo de parto cesárea indicou-se como um fator de maior risco para o óbito em relação ao parto normal.

A variação espacial do risco na região do município de Curitiba em 2004 para a mortalidade neonatal evidenciou-se constante ao longo do espaço controlando por fatores individuais, porém para a mortalidade pós-neonatal este risco apresentou-se aparentemente não constante, assim identificando algumas regiões com menor propensão ao óbito.

REFERÊNCIAS BIBLIOGRÁFICAS

CARVALHO, M. S.; SANTOS, R. S. **Análise de Dados Espaciais em Saúde Pública: métodos, problemas, perspectivas.** Caderno de Saúde Pública. Rio de Janeiro; 21(2): 361-378, mar/abr, 2005.

DRUCK, S; CARVALHO, M. S.; CAMARA, G.; MONTEIRO, A.V.M.. **Análise Espacial de Dados Geográficos**. Embrapa. Brasília, (ISBN: 85-7383-260-6), 2004.

GIOLO, S. R. **Análise de Regressão**. Curitiba, Universidade Federal do Paraná – Departamento de Estatística, 2003.

KOZU, K. T. ; GODINHO, L. T. ; MUNIZ ; M.V.F. ; CHIARIONI, P. **Mortalidade Infantil: Causas e Fatores de Risco**. **Um Estudo Bibliográfico**. Disponível em: <http://www.medstudents.com.br/original/original/mortinf/mortinf.htm> Acesso em 15 ago. 2006.

MENEGUETTE, M. ; ALVES, D. B. M. ; MONICO, J. F. G. . **Atenuação dos efeitos da ionosfera no posicionamento relativo gps de alta precisão utilizando a técnica dos mínimos quadrados penalizados.** Séries em Ciências Geodésicas, v. III, 2003.

MORAIS NETO, O. L.; BARROS, M. B. A.; MARTELLI, C. M. T.; SILVA, S.; CAVENAGHI, S. M.; SIQUEIRA JUNIOR, J. B.. **Diferenças no padrão de ocorrência da mortalidade neonatal e pós-neonatal no Município de Goiânia, Brasil, 1992-1996: análise espacial para identificação das áreas de risco**. Caderno de Saúde Pública, Rio de Janeiro, 17:1241-1250, set/out, 2001.

PÈREZ, F. L. **Critério AIC para a seleção de modelos.** Curitiba, Universidade Federal do Paraná – Departamento de Estatística, 2004.

RODRIGUES, M. A. S. **ÁRVORES DE CLASSIFICAÇÃO.** Açores, 2005. Monografia – Departamento de Matemática, Universidade dos Açores.

SHIMAKURA, S. E.; CARVALHO, M. S.; AERTS, D. R.G.C.; FLORES, R.. **Distribuição Espacial do Risco: Modelagem da Mortalidade em Porto Alegre-RS**. Caderno de Saúde Pública, Rio de Janeiro; 17(5): 1251-1261, set/out, 2001.

SOUZA, R. K. T.; GOTLIEB, S. L. D. **Probabilidade de morrer no primeiro ano de vida em área urbana da região Sul**. Rev. Saúde Pública, São Paulo, 27 (6): 445-454, 1.

WAHBA, G. **(Smoothing) Splines In Nonparametric Regression**. Madison, University Of Wisconsin - Department Of Statistics, 2000 Technical Report.

WOOD, S. N**. Generalized Additive Models: Introduction With R**. Boca Raton, Chapman & Hall/CRC, 2006.

BRAGA, L. P. V. **Introdução à Mineração de Dados.** Rio de Janeiro, E-papers Serviços Editoriais, 2005.

WOOD, S. N**. Tensor Product smooth interaction terms in Generalized Additive Mixed Models,** Glasgow, University of Glasgow - Department of Statistical, 2004.

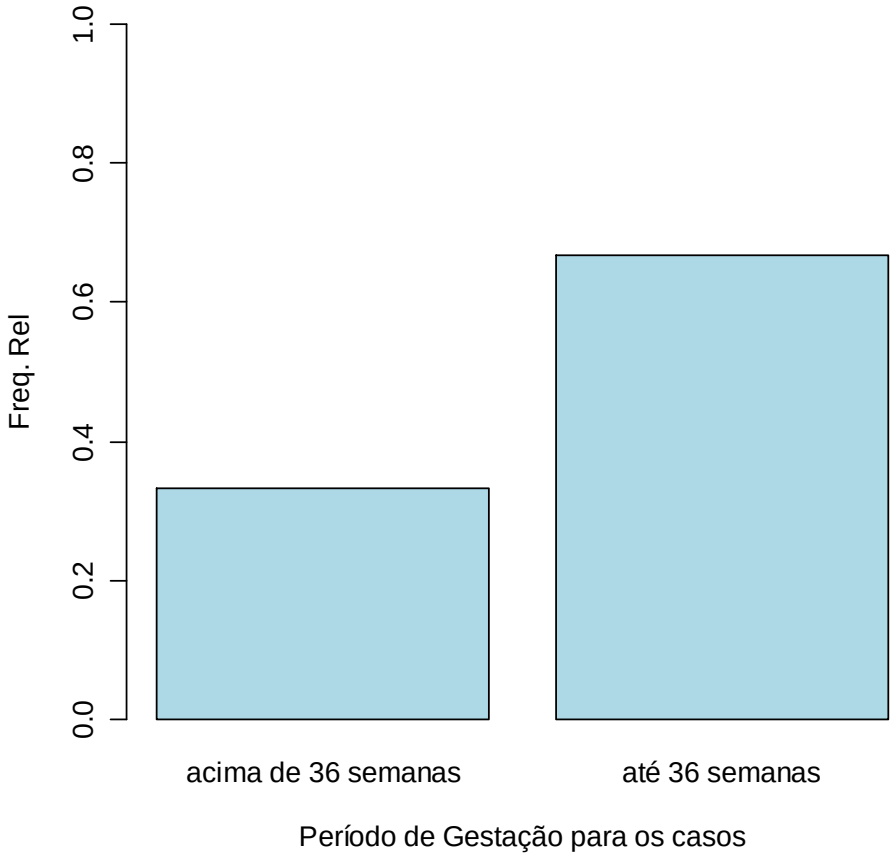
ANGOSS CORPORATION. **Knowledge Studio version 5.2**. Toronto, 2006. Data Mining, Windows XP.

ACL SERVICES LTD. **ACL version 8.4.1**. Vancouver, 2005. Business Assurance Analytics. Windows 2000.

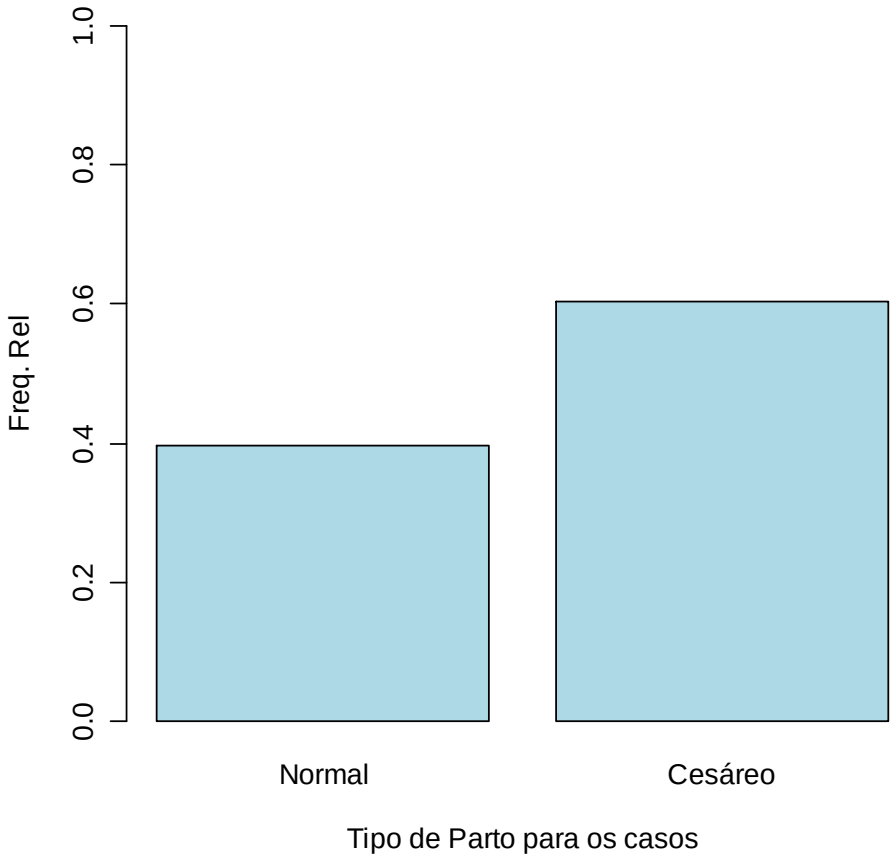
ANEXOS

Gráficos de Barras para as váriáveis Gestação, Tipo de Parto Sexo da criança e Escolaridade da mãe para a componente neonatal.

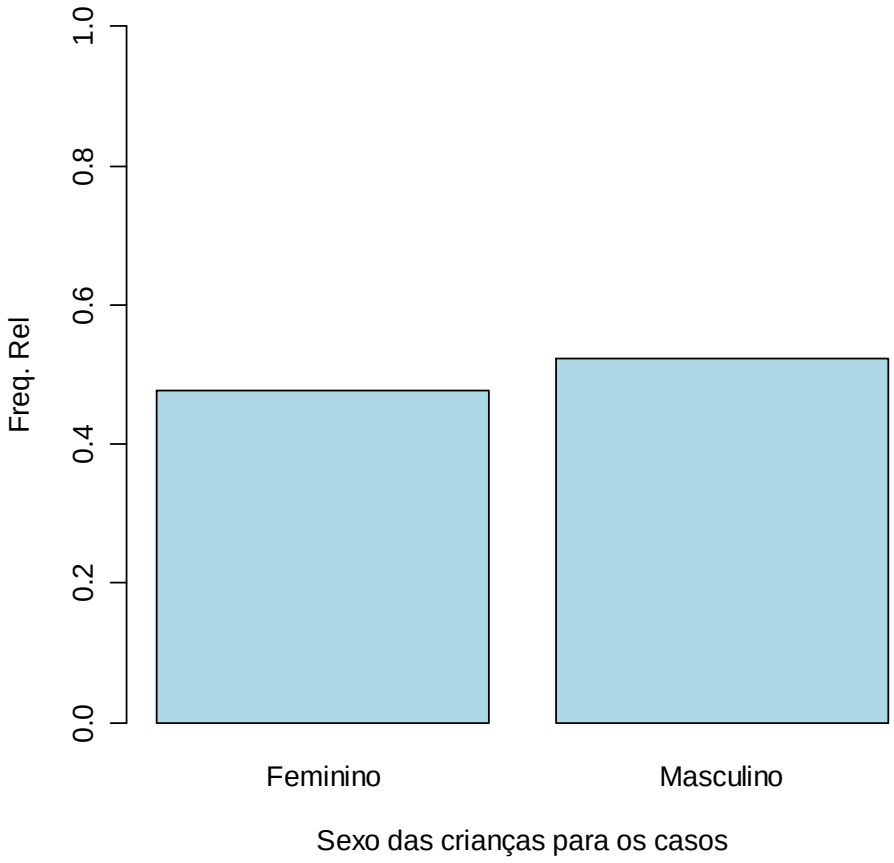
FREQUÊNCIA PARA GESTAÇÃO PARA OS CASOS NEONATAL



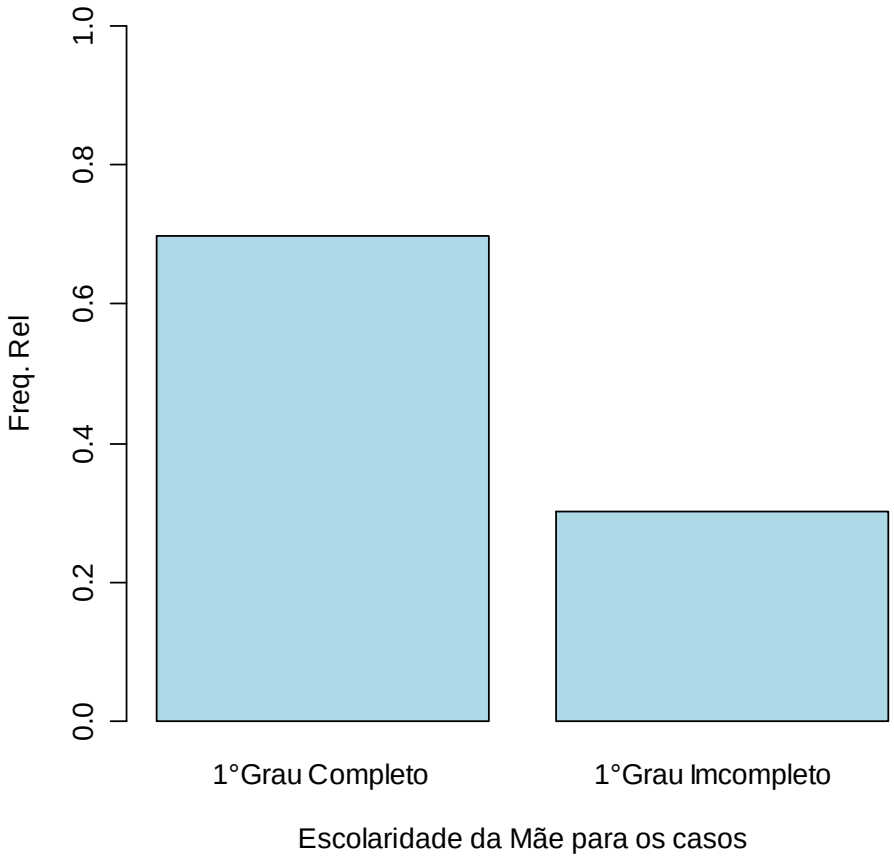
FREQUÊNCIA PARA TIPO DE PARTO PARA OS CASOS NEONATAL



FREQUÊNCIA PARA O SEXO DAS CRIANÇAS PARA OS CASOS NEONATAL

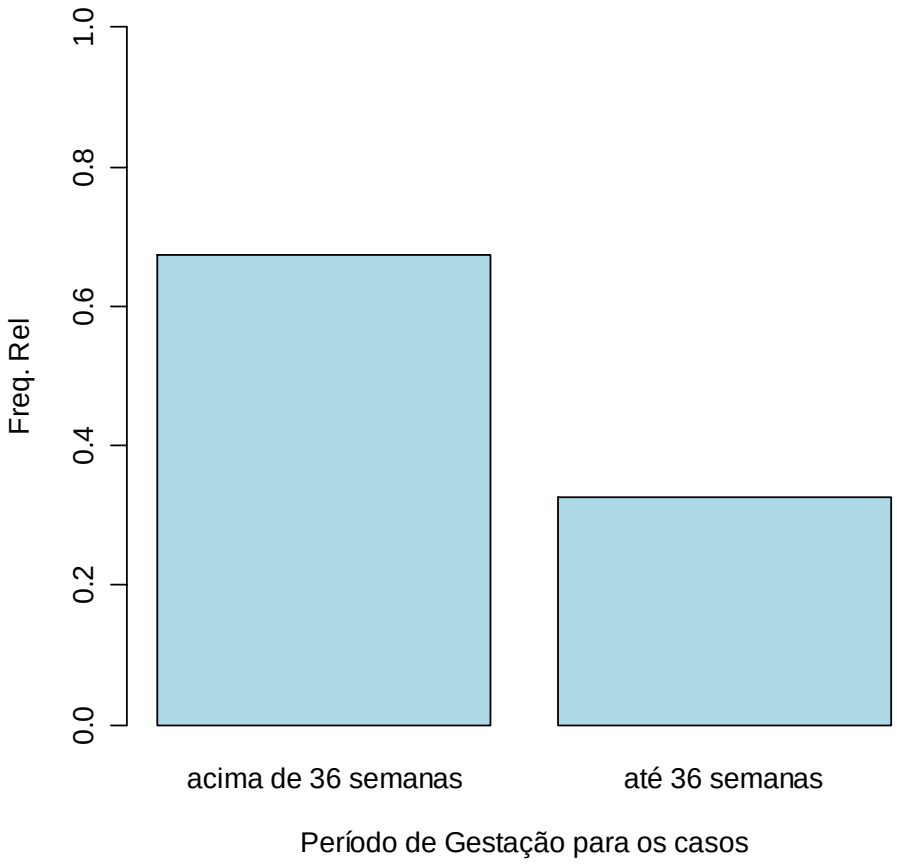


FREQUÊNCIA PARA ESCOLARIDADE DA MÃE PARA OS CASOS NEONATAL

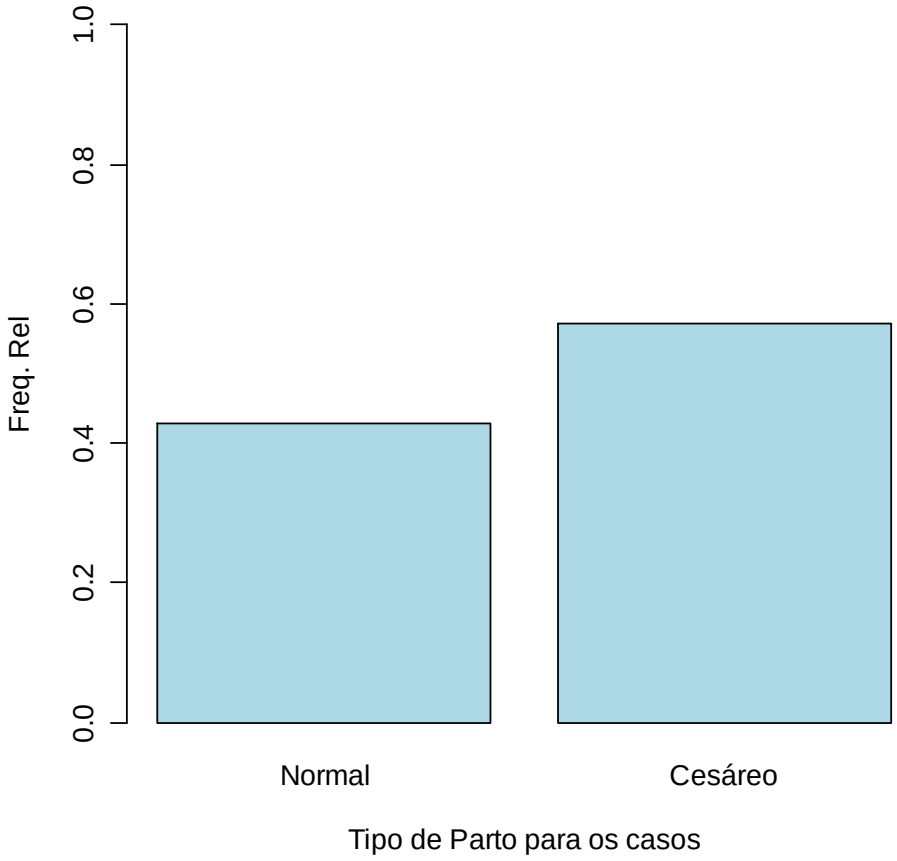


Gráficos de Barras para as váriáveis Gestação, Tipo de Parto, Sexo da criança e Escolaridade da mãe para a componente pós-neonatal.

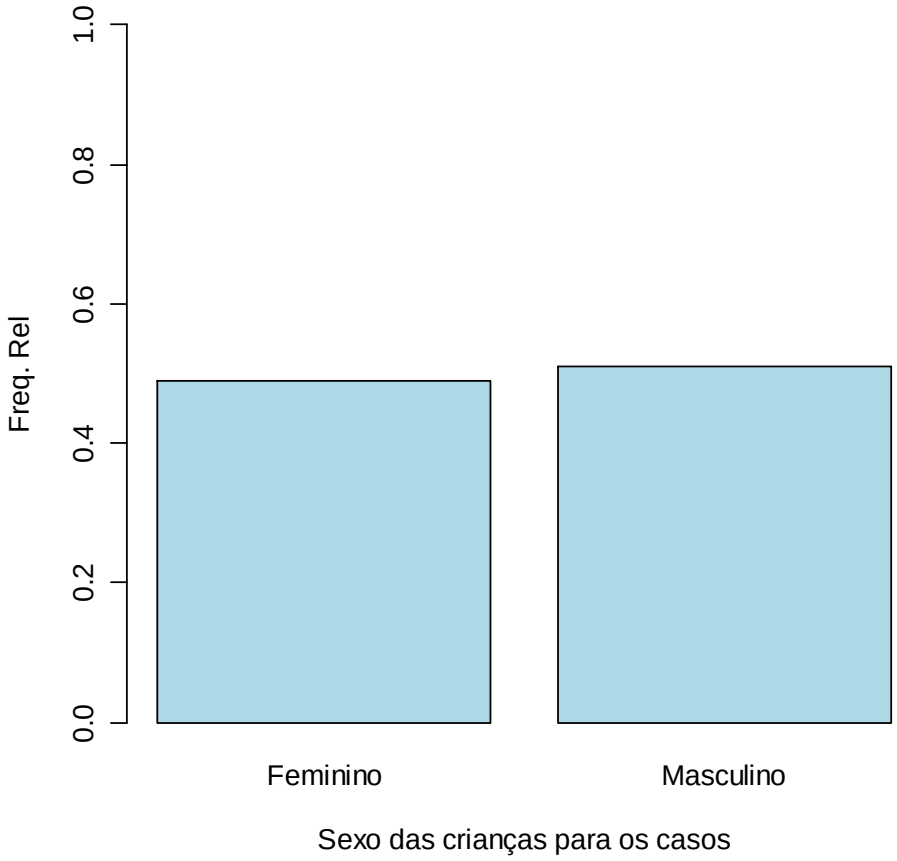
FREQUÊNCIA PARA GESTAÇÃO PARA OS CASOS PÓS-NEONATAL



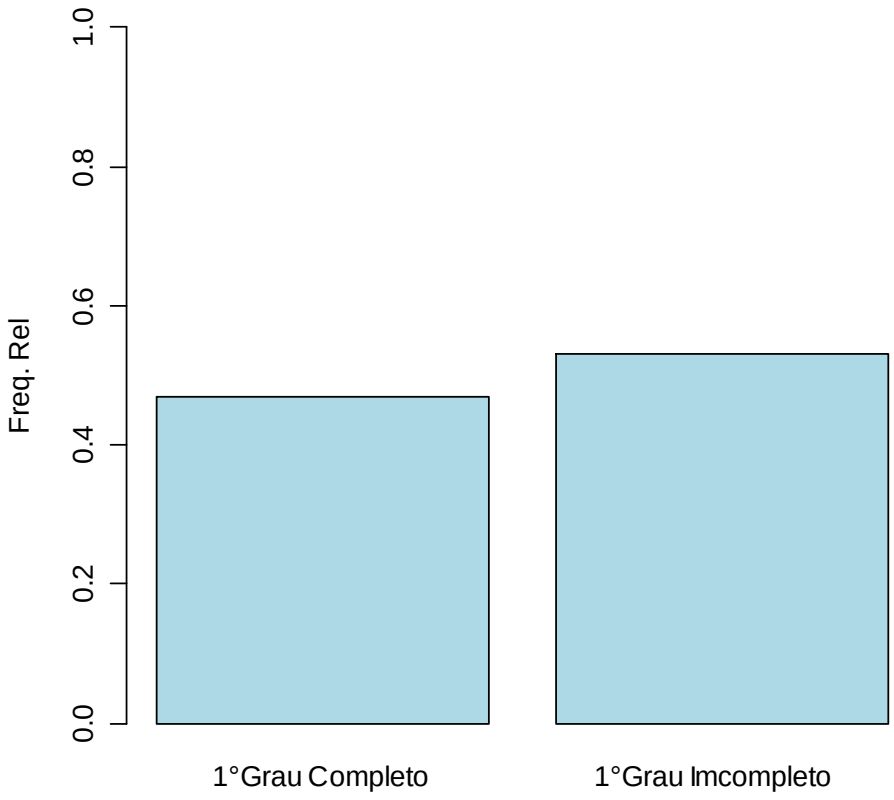
FREQUÊNCIA PARA TIPO DE PARTO OS CASOS PÓS- NEONATAL



FREQUÊNCIA PARA SEXO DA CRIANÇA PARA OS CASOS PÓS-NEONATAL



FREQUÊNCIA PARA ESCOLARIDADE DA MÃE PARA OS CASOS PÓS-NEONATAL



COMANDOS DO R USADOS NO TRABALHO:

Leitura do shapefile

Pacote necessário para ler:

```
require(shapefiles)
```

shape dos óbitos

```
sim=read.shapefile('c:/R/sim')
```

adiciona as coordenadas x e y no dbf

```
shapefile <- add.xy(sim)
```

```
names(shapefile)
```

Banco de Dados ‘sim’ com as Coordenadas

```
obito=(shapefile$dbf$dbf)
```

shape dos nascidos vivos

```
sinasc=read.shapefile('c:/R/sinasc')
```

adiciona as coordenadas x e y no dbf

```
shape=add.xy(sinasc)
```

Banco de Dados ‘sinasc’ com as Coordenadas

```
sinasc1=(shape$dbf$dbf)
```

Criar um arquivo para guardar os bancos de dados fora do 'R'

```
require(foreign)

write.dbf(obito, file = "c:/R/obito")

write.dbf(sinasc1, file = "c:/R/vivo")
```

Verificar a idade em dias, que a criança morreu

```
require(date)
sim=read.dbf('c:/R/obito.dbf')
a=as.date(as.character(DTNASC),order='dmy')
b=as.date(as.character(DTOBITO),order='dmy')
DIA=b-a
diaobito=data.frame(sim,DIA)
head(diaobito)
write.dbf(diaobito, "c:/R/dia_obito.dbf")
```

Seleção de controles para os bancos de dados:
neonatal e pos neonatal

Seleção dos controles para 'posneonatal'

```
require(foreign)
vivo=read.dbf('c:/R/vivo')
dim(vivo)
set.seed(569716937)
ind=(1:5480)
amostra=sample(ind,98)
cont_pos=vivo[amostra,]
write.dbf(cont_pos,'c:/R/cont_posneo')
```

Seleção dos controles para "neonatal"

```
set.seed(7)
ind=(1:5480)
amostra=sample(ind,274)
cont_neo=vivo[amostra,]
write.dbf(cont_neo,'c:/R/cont_neonatal')
```

Leitura dos bancos de dados:
neonatal do cont_neonatal

Leitura do banco de dados neonatal - neo_contr

```
bd_neo=read.table("c:/R/neo_contr.txt", h=T)
```

Leitura do banco de dados pós-neonatal - pos_contr

```
pos<-read.table("c:/R/pos_contr.txt",h=T)
```

Leitura do shape de Curitiba

```
require(shapefiles)

CURITIBA =read.shapefile('c:/R/div_municipal')
```

Transformação para plotar os obitos e os vivos dentro do mapa de Curitiba

```
require(maptools)

fylk.val <- shape2lines(CURITIBA)
xylims <- attr(fylk.val, "maplim")
plot(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="",col='#404040')
for (i in 1:length(fylk.val)) lines(fylk.val[[i]])
```

Plotar os óbitos e os controles dentro do mapa

Controles

```
plot(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="",col='#404040')
for (i in 1:length(fylk.val)) lines(fylk.val[[i]])

points(bd_neo$XCOORD[bd_neo$Y==0],bd_neo$YCOORD[bd_neo$Y==0],col='red',cex
= 0.1)
points(pos$XCOORD[pos$Y==0],pos$YCOORD[pos$Y==0],col=2)
```

Óbitos

```
plot(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])

points(pos$XCOORD[pos$Y==1],pos$YCOORD[pos$Y==1],col=2)
points(bd_neo$XCOORD[bd_neo$Y==1],bd_neo$YCOORD[bd_neo$Y==1],col=4)
```

Histogramas de freqüências da idade da mãe dividida por caso e controle

```
par(mfrow=c(1,2))
hist(IDADEMAE[Y==1],prob=T,ylim=c(0,0.07))
hist(IDADEMAE[Y==0],prob=T,ylim=c(0,0.07))
```

Sumarização das variáveis

Variáveis numéricas

Peso da criança

```
summary(pos$PESO[Y==0])
summary(pos$PESO[Y==1])
```

Histograma para o grupo NEONATAL

```
par(mfrow=c(1,2))
hist(bd_neo$PESO[bd_neo$Y==1],main="Histograma para \n o peso(g) - casos",
ylim=c(0,120),xlab='Peso(g)', ylab='Número de casos',xlim=c(0,4500))
hist(bd_neo$PESO[bd_neo$Y==0],main="Histograma para \n o peso(g) -
controles",
ylim=c(0,120),xlab='Peso(g)', ylab='Número de controles',xlim=c(0,4500))
```

Histograma para o grupo PÓS-NEONATAL

```
par(mfrow=c(1,2))
hist(pos$PESO[pos$Y==1],main="Histograma para \n o peso(g) - casos",
ylim=c(0,50),xlab='Peso(g)', ylab='Número de casos',xlim=c(0,4500))
hist(pos$PESO[pos$Y==0],main="Histograma para \n o peso(g) - controles",
ylim=c(0,50),xlab='Peso(g)', ylab='Número de controles',xlim=c(0,4500))
```

Idade da mãe

```
summary(pos$IDADEMAE[Y==0])
summary(pos$IDADEMAE[Y==1])
```

Histograma para o grupo NEONATAL

```
par(mfrow=c(1,2))
hist(bd_neo$IDADEMAE[bd_neo$Y==1],main="Histograma para \n a Idade da Mãe -
casos",
ylim=c(0,90),xlab='Idade', ylab='Número de casos',xlim=c(10,45))
hist(bd_neo$IDADEMAE[bd_neo$Y==0],main="Histograma para \n a Idade da Mãe -
controles",
ylim=c(0,90),xlab='Idade', ylab='Número de controles',xlim=c(10,45))
```

Histograma para o grupo PÓS-NEONATAL

```
par(mfrow=c(1,2))
hist(pos$IDADEMAE[pos$Y==1],main="Histograma para \n a Idade da Mãe -
casos",
ylim=c(0,40),xlab='Idade', ylab='Número de casos',xlim=c(10,45))
hist(pos$IDADEMAE[pos$Y==0],main="Histograma para \n a Idade da Mãe -
controles",
ylim=c(0,40),xlab='Idade', ylab='Número de controles',xlim=c(10,45))
```

Procura do melhor modelo ajustado

```
bd_neo=read.table("c:/R/dicot_neo_contr.txt", h=T)
summary(bd_neo)
names(bd_neo)
attach(bd_neo)
```

NEONATAL

```
require(mgcv)

ESTCIVMAE=as.factor(bd_neo$ESTCIVMAE)
```

Seleção do melhor modelo

Neonatal

```
mod0<-glm(Y~1, data=bd_neo)
step(mod0,
~ESCMAE*ESTCIVMAE*GESTACAO*GRAVIDEZ*IDADEMAE*PARTO*PESO*QTDFILMORT*QTDFILVI
VO*RACACOR*SEXO, family=binomial(link="logit"),direction=c("both"))
```

Modelo escolhido usando *Thin Plate*

```
models=gam(Y ~ PESO + GESTACAO + IDADEMAE + QTDFILVIVO + PARTO +
s(neo$XCOORD,neo$YCOORD, k=10, bs="tp"), family=binomial)

summary(models)
```

Mapa de Risco usando *Thin Plate*

```
par(mfrow=c(1,2))
```

```
vis.gam(models,plot.type="contour",view=c('neo$XCOORD','neo$YCOORD'),asp=1,
ylim=c(7160000,7200000),main="Usando Thin Plate", color='cm')

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```

Modelo escolhido usando *Tensor Product*

```
modelte=gam(Y ~ PESO + GESTACAO + IDADEMAE + QTDFILVIVO + PARTO
    + te(neo$XCOORD,neo$YCOORD, k=10), family=binomial)

summary(modelte)
```

Mapa de Risco usando *Tensor Product*

```
vis.gam(modelte,plot.type="contour",view=c('neo$XCOORD','neo$YCOORD'),asp=1
,ylim=c(7160000,7200000),main="Usando Tensor Product", color='cm')

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```

Mapa dos Resíduos

```
par(mfrow=c(1,2))
```

```
plot(models,residuals=T,asp=1,main="Usando Thin Plate")

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])


plot(modelte,residuals=T,asp=1,main="Usando Tensor Product")

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```

Pós-neonatal

```
pos<-read.table("c:/R/pos_contr.txt",h=T)
attach(pos)

ESTCIVMAE=as.factor(ESTCIVMAE)


require(mgcv)

mod0<-glm(Y~1, data=pos)
step(mod0,
~ESCMAE*ESTCIVMAE*GESTACAO*GRAVIDEZ*IDADEMAE*PARTO*PESO*QTDFILMORT*QTDFILVI
VO*RACACOR*SEXO, family=binomial(link="logit"),direction=c("both"))
```

Modelo escolhido usando *Thin Plate*

```
models=gam(Y ~ PESO + QTDFILVIVO + IDADEMAE +
s(pos$XCOORD,pos$YCOORD, k=10,bs="tp"), family=binomial)

summary(models)
```

Mapa de Risco usando *Thin Plate*

```
vis.gam(models,plot.type="contour",view=c('pos$XCOORD','pos$YCOORD'),asp=1,
ylim=c(7160000,7200000), main="Usando Thin Plate", color='cm')

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```

Modelo escolhido usando *Tensor Product*

```
modelte=gam(Y ~PESO + QTDFILVIVO + IDADEMAE  + QTDFILVIVO:IDADEMAE
           +te(pos$XCOORD,pos$YCOORD, k=10), family=binomial)

summary(modelte)
```

Mapa de Risco usando *Tensor Product*

```
vis.gam(modelte,plot.type="contour",view=c('pos$XCOORD','pos$YCOORD'),asp=1
,ylim=c(7160000,7200000), main="Usando Tensor Product", color='cm')

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```

Mapa dos Resíduos

```
par(mfrow=c(1,2))

plot(models,residuals=T,asp=1,main="Usando Thin Plate")

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])


plot(modelte,residuals=T,asp=1,main="Usando Tensor Product")

require(shapefiles)
curitiba=read.shapefile('c:/R/div_municipal')
require(maptools)
fylk <- shape2lines(curitiba)
xylims <- attr(fylk, "maplim")
lines(xylims$x, xylims$y, asp=1, type='n', xlab="", ylab="")
for (i in 1:length(fylk)) lines(fylk[[i]])
```