UMA PROPOSTA PARA ESTIMAÇÃO DA TAXA E DA SUBNOTIFICAÇÃO DE REGISTROS DE ESTUPRO DE VULNERÁVEL NO BRASIL

Natan Sant Anna Borges

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: James Dean Oliveira dos Santos Júnior

Manaus
Abril de 2018

UMA PROPOSTA PARA ESTIMAÇÃO DA TAXA E DA SUBNOTIFICAÇÃO DE REGISTROS DE ESTUPRO DE VULNERÁVEL NO BRASIL

Natan Sant Anna Borges

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinada por:

Prof. Plínio Lucas Dias Andrade, Dr.

Prof. Celso Rômulo Barbosa Cabral, Dr.

Prof. James Dean Oliveira dos Santos Júnior, Dr.

MANAUS, AM – BRASIL
ABRIL DE 2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Borges, Natan Sant' Anna
B732u Uma proposta para estimação da taxa e da subnotificação de registros de estupro de vulnerável no Brasil / Natan Sant' Anna Borges. 2018
43 f.: il.; 31 cm.

Orientador: James Dean Oliveira dos Santos Junior
Dissertação (Mestrado em Matemática - Estatística) - Universidade Federal do Amazonas.

1. Inferência Bayesiana. 2. Estupro de vulnerável . 3. Poisson multivariada. 4. Amostrador de Gibbs. I. Santos Junior, James Dean Oliveira dos II. Universidade Federal do Amazonas III. Título

"Todo mundo é um gênio. Mas, se você julgar um peixe por sua capacidade de subir em uma árvore, ela vai gastar toda a sua vida acreditando que ele é estúpido."

Albert Einstein

Agradecimentos

Agradeço a todos que, direta ou indiretamente, me apoiaram neste trabalho.

Ao professor James, meu orientador, pelos conhecimentos transmitidos, compreensão, atenção, incentivo e pela confiança depositada em mim.

A minha namorada Ranah, pelo companheirismo, força, incentivo, dedicação, carinho e compreensão.

Aos meus pais, Gerusa e Claudio, e meus irmãos, Thadeu, Lara e Enzo, que mesmo de longe foram a parte fundamental neste processo.

Aos amigos do Itapuca, do fundão do COMATH, da Estatística-UFF, companheiros de mestrado, da minha infância e demais amigos que fiz ao longo dos anos.

Aos colegas do IFAM e apoio nesta dupla jornada.

Aos professores do PPGM-UFAM pelo conhecimento transmitido.

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

UMA PROPOSTA PARA ESTIMAÇÃO DA TAXA E DA SUBNOTIFICAÇÃO DE REGISTROS DE ESTUPRO DE VULNERÁVEL NO BRASIL

Natan Sant Anna Borges

Abril/2018

Orientador: James Dean Oliveira dos Santos Júnior

Área de Concentração: Estatística

O estupro de vulnerável é um fenômeno que tem sofrido um considerado aumento em sua ocorrência, mas pouco se conhece sobre esses números devido ao alto índice de subnotificações referente a este tipo de crime. O objetivo deste estudo é estimar a taxa de estupro de vulnerável por meio da inferência bayesiana e técnicas de aumento de dados. A metodologia proposta utilizou os dados de estupro de vulnerável de algumas cidades do interior do estado do Amazonas referente ao período de 2010 a 2012. O emprego de um modelo Poisson multivariado mostrou-se eficaz, apresentando uma boa aderência aos dados supracitados.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

A PROPOSAL FOR ESTIMATION OF THE RATE AND UNDERREPORTING OF RAPE OF VULNERABLE IN BRAZIL

Natan Sant Anna Borges

April/2018

Advisor: James Dean Oliveira dos Santos Júnior

Research lines: Statistics

The rape of vulnerable people is a phenomenon that has experienced a considerable increase in its occurrence. However these numbers are not vastly known due to the high index of under-reportings related to this type of crime. This study aims to estimate the rate of rape of vulnerable people by means of the Bayesian inference and data augmentation techniques. The proposed methodology used the data of rape of vulnerable people of some countryside cities of Amazonas referring to the period of 2010 to 2012. Due to the good adherence to the aforementioned data, the deployment of a multivariate Poisson model showed its efficacy.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A legislação brasileira enquadra o crime de estupro dentre aqueles considerados crimes contra a liberdade sexual e, embora tenhamos penas que são consideradas elevadas, a taxa desse tipo de crime tem apresentado tendência de crescimento nos últimos anos [3]. Existem graves consequências do estupro, de curto e longo prazo, que se estendem no campo físico, psicológico e econômico. Além de lesões que a vítima pode sofrer nos órgãos genitais (principalmente nos casos envolvendo crianças), quando há o emprego de violência física, muitas vezes ocorrem também contusões e fraturas que, no limite, podem levar ao óbito da vítima [2].

Diferentes estudos apontam que o crime de estupro é aquele que apresenta a maior taxa de subnotificação. O 9ºAnuário Brasileiro de Segurança Pública, em 2014, estimou que apenas 35% dos casos de estupros são notificados. Em contrapartida [2] produziu um estudo que verificou que apenas 10% dos crimes de estupro são efetivamente notificados. Acredita-se que muitas vezes as ocorrências não são registradas por vergonha ou medo por parte das vítimas.

Resultados divulgados em [3] , destacaram o crescimento da taxa de estupros para cada 100 mil habitantes, no estado do Amazonas, que passou de 18,4 em 2015 para 23,2 em 2016, resultando um aumento de 26,1%. As estatísticas informadas correspondem apenas as ocorrências policiais registradas.

Destaca-se dentre os diversos tipos de estupros, o estupro de vulnerável, o qual será o foco da abordagem desta dissertação. A partir do artigo 217-A do Código Penal, entende-se estupro de vulnerável como a conjunção carnal ou a prática de qualquer ato libidinoso com pessoa menor de 14 anos. O consentimento da vítima, sua eventual experiência sexual anterior ou a existência de relacionamento amoroso entre o agente e a vítima não afastam a ocorrência do crime.

É razoável supor que a subnotificação seja maior para mulheres abaixo de 14 anos. Contudo, existem outros registros no Brasil que registram de maneira indireta este crime silencioso e que podem nos auxiliar a corrigir a taxa de estupro de vulneráveis. O primeiro é o registro de nascidos vivos (SINASC), no qual podemos estimar a

idade da mãe e o mês da concepção. O outro é o banco de registro de óbitos fetais (SIM). Este dois bancos de dados são gerenciados pelo SUS e estão disponíveis na plataforma DATASUS. Se em qualquer um destes bancos encontrarmos uma mãe com menos de 14 anos, podemos contar que houve um estupro de vulnerável, vale ressaltar que estes bancos referem-se somente à vítimas que engravidaram. Contudo, este registro só pode ser contado como não notificado se ele não aparecer nos dados das delegacias. Como a identidade do indivíduo é preservada tanto pelo SUS quanto pela Secretaria de Segurança Pública, faz-se importante a criação de mecanismos que nos permitam estimar a proporção de subnotificações através destes novos bancos de dados e por consequência corrigir a taxa de estupros de vulnerável.

1.1 Objetivo

O objetivo desta dissertação é construir um modelo probabilístico para estimar a taxa de estupro de vulnerável, corrigida a partir dos dados do SUS. Este modelo será aplicado nos dados de algumas cidades do Estado do Amazonas.

1.2 Organização do texto

No Capítulo 2 é feita uma revisão da literatura utilizada nos próximos capítulos, destacando-se as características da distribuição de Poisson multivariada que será proposta como modelo para os dados aumentados. Em seguida, foi apresentado o conceito de inferência bayesiana e o método de aproximação MCMC usado para a obtenção das distribuições marginais a posteriori. O capítulo termina apresentando a medida Deviance Information Criterion (DIC), utilizada para seleção de modelos. No Capítulo 3 é descrita de forma resumida como foram obtidos os dados. Além disso, relata-se como foi realizada a estimativa das idades das vítimas. O Capítulo 4 descreve a estrutura dos modelos para o ajuste dos dados de estupro de vulnerável. Em adição, aborda-se como a inferência bayesiana é realizada usando o MCMC. O Capítulo 5 apresenta a aplicação da modelagem proposta. Primeiramente, é apresentada uma análise exploratória dos dados. Em seguida, são apresentados diversos modelos e os seus resultados utilizando a metodologia descrita no capítulo anterior. No Capítulo 6 discute as vantagens do modelo escolhido para estimar a taxa de estupro de vulnerável.

Capítulo 2

Revisão de Literatura

Neste capítulo é apresentada uma breve revisão sobre o modelo Poisson multivariado, que será proposto para modelar os dados em estudo. Também são apresentados os principais conceitos de inferência bayesiana, amostrador de Gibbs e o critério DIC para seleção de modelos.

2.1 Poisson Multivariada

Um modelo multivariado discreto recebe o nome de Poisson multivariado quando suas distribuições marginais possuem distribuição Poisson. Este modelo é uma alternativa para tratar dados discretos multivariados. Neste trabalho será utilizado o modelo proposto por [11], que apresenta a vantagem de incluir a informação da correlação proveniente dos dados no processo de modelagem diretamente na distribuição de probabilidade dos dados e não em outros níveis da hierarquia do modelo.

No caso trivariado, o vetor $\mathbf{Y}= (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$ com $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_{12}, \lambda_{13}, \lambda_{23}, \lambda_{123})$, tal que:

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123} \tag{2.1}$$

$$Y_2 = X_2 + X_{12} + X_{23} + X_{123} \tag{2.2}$$

$$Y_3 = X_3 + X_{13} + X_{23} + X_{123} \tag{2.3}$$

onde as $X_i's$ são variáveis aleatórias independentes com distribuição de Poisson univariada de parâmetros λ_i ,$i \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\}, \{23\}, \{123\})$. Desta forma, a função de probabilidade é dada por:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{x} \in C} k(\mathbf{x}, \mathbf{y}) e^{-\sum_i \lambda_i} \prod_{j=1}^{3} \lambda_j^{y_j - z_j(\mathbf{x})} \lambda_{12}^{x_{12}} \lambda_{13}^{x_{13}} \lambda_{23}^{x_{23}} \lambda_{123}^{x_{123}}, \tag{2.4}$$

onde $\mathbf{x} = (x_{12}, x_{13}, x_{23}, x_{123})$ e $z_j(\mathbf{x})$ é a soma de todos os elementos em $\{x_{12}, x_{13}, x_{23}, x_{123}\}$ que contém o número j no índice e

$$k(\mathbf{x}, \mathbf{y})^{-1} = x_{12}!x_{13}!x_{23}!x_{123}! \prod_{j=1}^{3} (y_j - z_j(\mathbf{z}))!, \tag{2.5}$$

onde a soma é dada por $C \subset \mathcal{N}^4$ que é definido como:

$$\begin{aligned} C = [(x_{12}, x_{13}, x_{23}, x_{123}) \in \mathcal{N}^4 : \{x_{12} + x_{13} + x_{123} \leq y_1\} \cap \{x_{12} + x_{23} + x_{123} \leq y_2\} \\ \cap \{x_{12} + x_{23} + x_{123} \leq y_3\}]. \end{aligned} \tag{2.6}$$

Temos que $\mathbf{Y} = (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$, possui vetor de médias igual a

$$E(\mathbf{Y}) = \begin{pmatrix} \lambda_1 + \lambda_{12} + \lambda_{13} + \lambda_{123} \\ \lambda_2 + \lambda_{12} + \lambda_{23} + \lambda_{123} \\ \lambda_3 + \lambda_{13} + \lambda_{23} + \lambda_{123} \end{pmatrix} \tag{2.7}$$

A principal restrição deste modelo é a exclusão de correlações negativas, devido ao fato de que a matriz de covariância é dada por:

$$\Sigma = \begin{bmatrix} \lambda_1 + \lambda_{12} + \lambda_{13} + \lambda_{123} & & \\ \lambda_{12} + \lambda_{123} & \lambda_2 + \lambda_{12} + \lambda_{23} + \lambda_{123} & \\ \lambda_{13} + \lambda_{123} & \lambda_{23} + \lambda_{123} & \lambda_3 + \lambda_{13} + \lambda_{23} + \lambda_{123} \end{bmatrix} \tag{2.8}$$

Assim a matriz de correlação será:

$$\rho = \begin{bmatrix} 1 & & \\ \frac{\lambda_{12}+\lambda_{123}}{\sqrt{\lambda_2+\lambda_{12}+\lambda_{23}+\lambda_{123}}\sqrt{\lambda_1+\lambda_{12}+\lambda_{13}+\lambda_{123}}} & 1 & \\ \frac{\lambda_{13}+\lambda_{123}}{\sqrt{\lambda_3+\lambda_{13}+\lambda_{23}+\lambda_{123}}\sqrt{\lambda_1+\lambda_{12}+\lambda_{13}+\lambda_{123}}} & \frac{\lambda_{23}+\lambda_{123}}{\sqrt{\lambda_3+\lambda_{13}+\lambda_{23}+\lambda_{123}}\sqrt{\lambda_2+\lambda_{12}+\lambda_{23}+\lambda_{123}}} & 1 \end{bmatrix} \tag{2.9}$$

Ou seja, $\lambda_{12}, \lambda_{13}, \lambda_{23}$ e λ_{123} são medidas de dependência entre as três variáveis aleatórias Y_1, Y_2 e Y_3. Se $\lambda_{12} = \lambda_{13} = \lambda_{23} = \lambda_{123} = 0$ então a distribuição trivariada acima é reduzida ao caso do produto de três distribuições Poisson independentes.

2.2 Inferência bayesiana

Esta seção apresenta, de maneira resumida, o procedimento utilizado na estimação dos parâmetros do modelo proposto. Detalhes mais específicos sobre inferência bayesiana podem ser vistos em [14].

A inferência bayesiana, ao contrário da inferência clássica, leva em conta o conceito de probabilidade subjetiva, que mede o grau de incerteza sobre os parâmetros do modelo. Assim, o procedimento de estimação consiste em atualizar as informações provenientes dos dados, resumida na função de verossimilhança, com a incerteza sobre os parâmetros, através de uma distribuição a priori, $\pi(\theta)$. O resultado deste procedimento é uma distribuição de probabilidade, $\pi(\theta \mid \mathbf{x})$ dita distribuição a posteriori, definida por:

$$\pi(\theta \mid \mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x} \mid \theta)}{f(\mathbf{x})} = \frac{\pi(\theta) f(\mathbf{x} \mid \theta)}{\int_\Theta \pi(\theta) f(\mathbf{x} \mid \theta) d\theta} \tag{2.10}$$

onde, $f(\mathbf{x} \mid \theta)$ denota a verossimilhança de θ dada a amostra $\mathbf{x}$. Além disso, vale notar que o denominador da equação anterior é uma constante em θ, ou seja, podemos escrever

$$\pi(\theta \mid \mathbf{x}) \propto \pi(\theta) f(\mathbf{x} \mid \theta). \tag{2.11}$$

No contexto de estimação dos parâmetros, seja $\delta(\mathbf{X})$ um estimador para θ com distribuição a posteriori $\pi(\theta \mid \mathbf{x})$. Seja $l(\theta, a)$ a função de perda associada à estimativa a de θ quando o verdadeiro valor do parâmetro é θ, de modo que quanto maior a distância entre a e θ, maior é a perda. Para cada estimativa a, a perda esperada a posteriori é definida por $E[l(\theta, a) \mid \mathbf{x}] = \int_\Theta l(\theta, a)\pi(\theta \mid \mathbf{x}) d\theta$. Seja $\delta(\mathbf{x})^*$ o valor de a para qual a perda esperada é mínima para todos os valores possíveis de $\mathbf{x}$. A função $\delta(\mathbf{X})^*$ é chamada de estimador de Bayes para θ e $\delta(\mathbf{x})^*$ a estimativa de Bayes tal que, para todos os valores possíveis de $\mathbf{x}$

$$E[l(\theta, \delta^*(\mathbf{x})) \mid \mathbf{x}] = \arg\min_a E[l(\theta, a) \mid \mathbf{x}]. \tag{2.12}$$

Diferentes funções de perda podem ser consideradas. Quando $l(\theta, a) = (\theta - a)^2$ (função de perda quadrática), o estimador de Bayes é a esperança da distribuição a posteriori para θ [14]. Tal estimador será utilizado neste trabalho.

Ao estimar pontualmente o parâmetro θ resume-se a informação da distribuição a posteriori em apenas um único valor. Uma maneira de representar as incertezas

sobre a estimativa pontual é a utilização de intervalos, que na inferência bayesiana são chamados de intervalos de credibilidade $100(1-\alpha)\%$, podendo ser obtidos a partir da distribuição a posteriori. Seja o parâmetro desconhecido $\theta \in \Theta$. Uma região $C \subset \Theta$ é um Intervalo de Credibilidade de $100(1-\alpha)\%$ para θ se:

$$P(\theta \in C \mid \mathbf{x}) \geq 1 - \alpha. \tag{2.13}$$

Quanto maior o intervalo, mais dispersa será a distribuição de θ, já ao contrário, quanto menor o intervalo, mais concentrada será a distribuição de θ.

2.3 Método de simulação Monte Carlo via Cadeias de Markov (MCMC)

Em geral, a distribuição a posteriori para θ é intratável analiticamente, inviabilizando a obtenção de amostras dessa distribuição por meio de técnicas como o Método da Inversão ou o algoritmo de Aceitação/Rejeição. Para simular dessas distribuições podemos utilizar métodos de simulação estocástica conhecidos como Monte Carlo via Cadeias de Markov (MCMC) [5].

Dentre as diversas abordagens MCMC existentes, neste estudo, foi utilizado o amostrador de Gibbs, introduzido originalmente por [8], e utilizado na inferência bayesiana após comparação com esquemas de simulação estocástica feitas por [6].

2.3.1 O amostrador de Gibbs

O método consiste em retirar, sucessivamente, amostras da distribuição da i-ésima componente do vetor de parâmetros, sendo os outros valores dos parâmetros no modelo considerados conhecidos ou fixos. Assim, o amostrador de Gibbs é um esquema amostral de uma cadeia de Markov cujo núcleo de transição é formado pelas condicionais completas e cuja distribuição estacionária é a marginal a posteriori dos parâmetros.

Para descrever o algoritmo, temos como distribuição de interesse a distribuição $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, onde $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$. Cada θ_i pode ser um escalar ou um vetor. Considere também que todas as condicionais completas $\pi(\theta_i \mid \theta_{-i}, \mathbf{x})$, $i = 1, \cdots, p$, em que θ_{-i} é um escalar ou um vetor formado por todos os parâmetros menos o θ_i, estejam disponíveis e que é possível gerar amostras de cada uma delas. Portanto, o algoritmo é dado por:

1. Inicialize $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_p^{(0)})$, e faça $t = 1$;

2. Gere um valor para $\boldsymbol{\theta}^{(t)}$, componente a componente, da forma

$$\theta_1^{(t)} \sim \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, \mathbf{x}) \tag{2.14}$$
$$\theta_2^{(t)} \sim \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, \mathbf{x}) \tag{2.15}$$
$$\vdots \tag{2.16}$$
$$\theta_i^{(t)} \sim \pi(\theta_i|\theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_p^{(t-1)}, \mathbf{x}) \tag{2.17}$$
$$\vdots \tag{2.18}$$
$$\theta_p^{(t)} \sim \pi(\theta_p|\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \cdots, \theta_{p-1}^{(t)}, \mathbf{x}); \tag{2.19}$$

3. Atualize o contador $t = t + 1$ e volte para (2) repetindo o procedimento até as cadeias geradas estabilizarem.

Depois de M iterações do processo, obtemos $(\theta_1^{(M)}, \theta_2^{(M)}, \cdots, \theta_p^{(M)})$.Foi demonstrado em [8] que para M suficientemente grande, a distribuição conjunta de $(\theta_1^{(M)}, \theta_2^{(M)}, \cdots, \theta_p^{(M)}) \mid \mathbf{x}$ converge em distribuição para $\pi(\theta_1, \theta_2, \cdots, \theta_p \mid \mathbf{x})$.

Para algoritmos MCMC é importante verificar a convergência das cadeias, pois neste caso podemos considerar que a amostra obtida vêm da distribuição alvo. Algumas técnicas para avaliar a convergência são discutidas em [1], [7] e [10]. Neste trabalho foi utilizada a técnica proposta por [7], que sugere como critério de convergência a utilização de várias cadeias em paralelo, começando de valores iniciais distintos. Para cada parâmetro de interesse, compara-se a variabilidade dentro e entre as cadeias amostradas, e uma vez que a estacionaridade tenha sido atingida, consideram-se as realizações como uma amostra aleatória da distribuição desejada. A convergência é monitorada através do fator de redução de escala potencial, expresso por $\hat{R}$. Se o valor de $\hat{R}$ for grande, será necessário considerar mais iterações para obter uma melhor estimativa dos parâmetros. Quando $\hat{R} \approx 1$, há evidências de que ocorreu a convergência.

2.4 Critério de seleção de modelos

Uma medida de comparação de modelos sob o enfoque clássico é a deviance, $D(\boldsymbol{\theta})$, definida como:

$$D(\theta) = -2\log(f(\mathbf{x} \mid \boldsymbol{\theta})). \tag{2.20}$$

Já no enfoque bayesiano,[18] propõe o DIC (*Deviance Information Criterion*), medida construída a partir de dois componentes: um termo mede a aderência e outro termo de penalidade pelo número efetivo de parâmetros

O primeiro termo, a medida de aderência, é definida pela esperança a posteriori da deviance

$$\bar{D} = E_{\boldsymbol{\theta}|\mathbf{x}}[D(\boldsymbol{\theta})]. \tag{2.21}$$

O segundo componente mede a complexidade do modelo através do número efetivo de parâmetros definido como a diferença entre a média posterior da deviance e a deviance avaliada na média posterior de $\boldsymbol{\theta}$.

$$pd = E_{\boldsymbol{\theta}|\mathbf{x}}[D(\boldsymbol{\theta})] - D[E_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta})] = \bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{2.22}$$

Então, o DIC é definido como a soma de ambos os componentes

$$DIC = \bar{D} + pd = 2\bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{2.23}$$

Usualmente, as quantidades $\bar{D}$ e $D(\bar{\theta})$ podem ser estimadas respectivamente, por

$$\frac{1}{M}\sum_{t=1}^{M} D(\theta^{(t)}) \quad \text{e} \quad D\left(\frac{1}{M}\sum_{t=1}^{M}\theta^{(t)}\right) \tag{2.24}$$

em que $\theta^{(t)}$, é o t - ésimo valor da amostra MCMC simulada válida, isto é, a amostra MCMC obtida depois de descartar o *burn-in* (número de iterações descartadas a fim de evitar a influência dos valores iniciais) e a reamostragem entre as cadeias.

Capítulo 3

Descrição do banco de dados

Os dados utilizados neste estudo foram obtidos por meio de *download* dos arquivos do Sistema de Informação sobre Mortalidade (SIM), e do Sistema de Informação sobre Nascidos Vivos (SINASC), disponíveis no site do Ministério da Saúde (www.datasus.gov.br) e dos boletins de ocorrência das delegacias (cedidos pelo Observatório de Violência de Gênero do Amazonas (Programa Institucional da UFAM)). Foram analisadas as cidades de Amaturá, Atalaia do Norte, Barreirinha, Benjamin Constant, Boa Vista do Ramos, Maués, Nhamundá, Santo Antônio do Içá, São Paulo de Olivença, Tabatinga e Tonantins, localizadas no interior do estado do Amazonas, entre os anos de 2010 a 2012.

Segundo o artigo 217-A do Código Penal, comete crime de estupro de vulnerável todo aquele que tiver conjunção carnal ou praticar qualquer outro ato libidinoso com pessoa menor de 14 anos. Logo, foram selecionados das bases de dados apenas as observações que apresentavam a idade no ato ocorrido inferior a 14 anos. Como apenas a base dados dos boletins de ocorrencia apresentou esta informação, para as demais bases utilizadas neste estudo foi necessário estimar a idade no ato da concepção atráves das váriavéis já existentes nas bases.

Na base de dados do SINASC foi considerado a variável “data da última menstruação” como o mês do ato da concepção, para então ser calculada a idade da vítima através da subtração da variável “data da última menstruação” pela variável “data de nascimento da mãe”. Seguindo essa operação, a estimtaiva pode apresentar um erro máximo de um mês da verdadeira data do ato de ocorrência.

Na base de dados do SIM foi utilizada a variável “idade da mãe"como a informação sobre a idade da vítima. Essa escolha acarretou a chance de um erro máximo de nove meses acrescentados na verdadeira idade do ato da concepção. Desta forma, foram desconsideradas as mulheres cuja a idade registrada fosse maior que 14 anos e 9 meses.

Neste trabalho foi verificado como se comporta as notificações entre as vítimas com idade inferior a 14, 13 e 12 anos. As Tabelas 3.1, 3.2 e 3.3 apresentam o

quantitativo de estupro de vulnerável nos três bancos analisados.

Tabela 3.1: Notificações de violência de vítimas abaixo de 14 anos

Cidades	Delegacia	SINASC	SIM
Amatura	2	9	0
Atalia do Norte	5	23	0
Barreirinha	9	23	0
Benjamin Constant	2	34	0
Boa Vista Do Ramos	6	16	0
Máues	12	63	1
Nhamundá	8	15	0
Santo Antônio do Içá	5	23	1
São Paulo de Olivença	4	19	0
Tabatinga	6	36	0
Tonantins	1	9	0
Total	**60**	**270**	**2**

Tabela 3.2: Notificações de violência de vítimas abaixo de 13 anos

Cidades	Delegacia	SINASC	SIM
Amatura	0	3	0
Atalia do Norte	3	5	0
Barreirinha	8	8	0
Benjamin Constant	1	7	0
Boa Vista Do Ramos	5	4	0
Máues	10	13	1
Nhamundá	7	4	0
Santo Antônio do Içá	3	6	1
São Paulo de Olivença	3	8	0
Tabatinga	5	6	0
Tonantins	1	1	0
Total	**46**	**65**	**2**

Tabela 3.3: Notificações de violência de vítimas abaixo de 12 anos

Cidades	Delegacia	SINASC	SIM
Amatura	0	0	0
Atalia do Norte	0	2	0
Barreirinha	5	1	0
Benjamin Constant	1	3	0
Boa Vista Do Ramos	4	0	0
Máues	4	3	1
Nhamundá	4	1	0
Santo Antônio do Içá	3	0	0
São Paulo de Olivença	3	3	0
Tabatinga	2	1	0
Tonantins	0	1	0
Total	**26**	**15**	**1**

Figura 3.1: Distribuição dos registros de estupro de vulnerável das vítimas abaixo de 14 anos

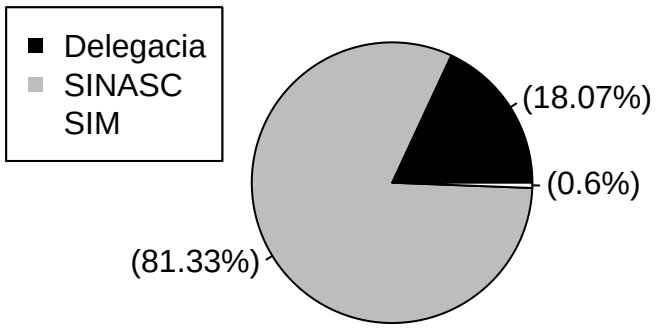


Figura 3.2: Distribuição dos registros de estupro de vulnerável das vítimas abaixo de 13 anos

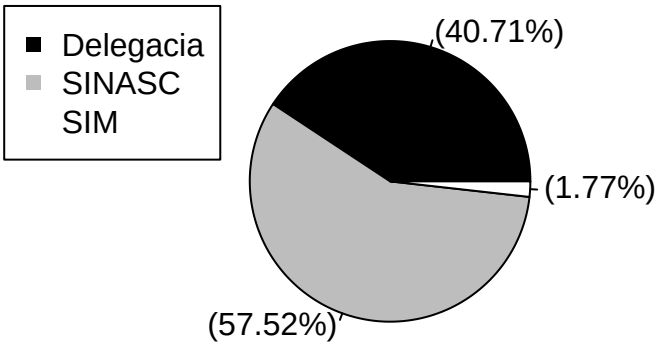
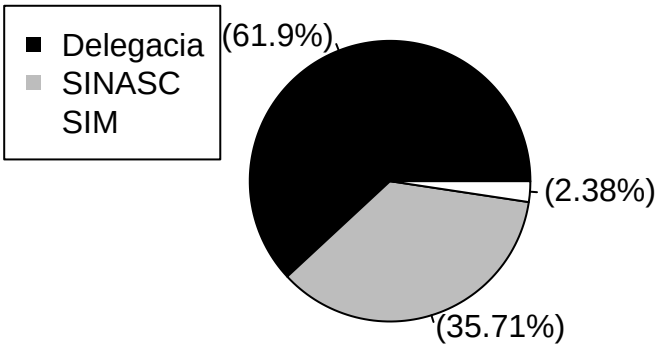


Figura 3.3: Distribuição dos registros de estupro de vulnerável das vítimas abaixo de 12 anos



As Figura 3.1, 3.2 e 3.3 mostram o aumento da proporção de casos de estupro de vulnerável registrados nas delegacias a medida que a idade da vítima diminui.

As Tabelas 3.4, 3.5 e 3.6 apresentam as matrizes de correlações amostral entre os dados da delegacia, SINASC e SIM nos três bancos analisados. Estas correlações foram calculadas como pré-requisito para modelar a distribuição 3-Poisson. Note que apenas uma das correlações foi negativa (Tabela 3.6), mas com um valor muito próximo de zero.

Tabela 3.4: Matriz de correlação amostral dos registros de estupro de vulnerável das vítimas abaixo de 14 anos

	Delegacia	SINASC	SIM
Delegacia	1.000	0.479	0.338
SINASC	0.479	1.000	0.452
SIM	0.338	0.452	1.000

Tabela 3.5: Matriz de correlação amostral dos registros de estupro de vulnerável das vítimas abaixo de 13 anos

	Delegacia	SINASC	SIM
Delegacia	1.000	0.520	0.302
SINASC	0.520	1.000	0.413
SIM	0.302	0.413	1.000

Tabela 3.6: Matriz de correlação amostral dos registros de estupro de vulnerável das vítimas abaixo de 12 anos

	Delegacia	SINASC	SIM
Delegacia	1.000	-0.015	0.302
SINASC	-0.015	1.000	0.417
SIM	0.306	0.417	1.000

Capítulo 4

Estrutura do modelo

Neste capítulo apresentamos a análise Bayesiana aplicada ao modelo de 3-Poisson para tratar sobre os dados referentes ao estupro de vulnerável provenientes da base de dados do SINAC, SIM e dos boletins de ocorrência.

4.1 Especificação para os modelos utilizados

O fato do modelo apresentado na Seção 2.1 ser usado para dados de contagem o torna uma razoável opção na modelagem proposta aqui, tendo em vista que denotamos Y_1 como o número de registros na delegacia, Y_2 como o número de registros no SINASC e Y_3 como o número de registros no SIM. O objetivo é construir um modelo que consiga explicar os dados através de estimativas para os parâmetros de interesse com uma pequena incerteza associada à estimação. Para isso, foram proposto quatro modelos multivariados.

O primeiro modelo, leva em consideração que existem registros em comum entre as base de dados do SINASC e da delegacia de polícia e também que a base de dados do SIM também apresenta registros em comum com a base da delegacia de polícia. As bases do SINASC e do SIM não apresentaram registros em comum, devido as bases tratarem de contagens distintas. Foi utilizado o lema apresentado abaixo, onde [11] demonstra a condição de independência entre as variáveis analisadas.

Lema 4.1.1 *Assume-se* $\boldsymbol{Y}= (Y_1, Y_2, Y_3)'$ *possui distribuição 3-Poisson onde* Y_1, Y_2 *e* Y_1, Y_3 *são variáveis dependentes e possuem distribuição de Poisson bivariada e assumindo que* $\lambda_{123} = \lambda_{23} = 0$ *então* Y_2, Y_3 *são variáveis independentes com distribuição de Poisson*

Uma vez assumido o Lema 4.1.1, obtém-se o modelo 1:

$$\begin{aligned} Y_1 &= X_1 + X_{12} + X_{13} \\ Y_2 &= X_2 + X_{12} \\ Y_3 &= X_3 + X_{13} \end{aligned} \tag{4.1}$$

Em que:
X_1 é o número de registros exclusivos da Delegacia
X_2 é o número de registros exclusivos do SINASC
X_3 é o número de registros exclusivos do SIM
X_{12} é o número de registros simultâneos na Delegacia e no SINASC
X_{13} é o número de registros simultâneos na Delegacia e no SIM

Além disso, definimos λ como a taxa esperada de estupro de vulnerável registradas; p_1 como a probabilidade do registro ser exclusivo da Delegacia ; p_2 como a probabilidade do registro ser exclusivo do SINASC; p_3 como a probabilidade do registro ser exclusivo do SIM; p_{12} como a probabilidade do registro ser simultâneo na Delegacia e no SINASC, e p_{13} como a probabilidade do registro ser simultâneo na na Delegacia e no SIM, em que $p_1 + p_2 + p_3 + p_{12} + p_{13} = 1$ e E como o total da população da cidade em questão (foi utilizada para os três anos o total de mulheres com idades entre 10 e 14 anos dados pelo Censo de 2010).

Logo as distribuições X_j possuem distribuição de Poisson com média $\lambda p_j E$ para $j \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\})$. Consequentemente $\mathbf{Y} = (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$ com $\Lambda = (E(\lambda p_1), E(\lambda p_2), E(\lambda p_3), E(\lambda p_{12}), E(\lambda p_{13}))$, com função de probabilidade conjunta definida por:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{x} \in C} k(\mathbf{x}, \mathbf{y}) e^{-E\lambda} \prod_{j=1}^{3} [E(\lambda p_j)]^{y_j - z_j(\mathbf{x})} E(\lambda p_{12})^{x_{12}} E(\lambda p_{13})^{x_{13}} \tag{4.2}$$

onde $\mathbf{x} = (x_{12}, x_{13})$ e $z_j(\mathbf{x})$ é a soma de todos os elementos em $\{x_{12}, x_{13}\}$ que contém o número j no índice e

$$k(\mathbf{x}, \mathbf{y})^{-1} = x_{12}! x_{13}! \prod_{j=1}^{3} (y_j - z_j(\mathbf{z}))! \tag{4.3}$$

onde a soma é dada por $C \subset \mathcal{N}^2$ que é definido como:

$$C = [(x_{12}, x_{13} \in \mathcal{N}^2 : \{x_{12} + x_{13} \leq y_1\} \cap \{x_{12} \leq y_2\} \cap \{x_{13} \leq y_3\}].$$

O modelo 2 leva em consideração que só existem notificações em comum entre a base de dados dos boletins de ocorrência e o SIM, e que não existe probabilidade de ocorrer registro em comum entre as bases de boletins de ocorrência e SINASC , ou seja, $p_{12} = 0$. Com isso, obtém-se o modelo 2:

$$\begin{aligned} Y_1 &= X_1 + X_{13}, \\ Y_2 &= X_2, \\ Y_3 &= X_3 + X_{13}. \end{aligned} \tag{4.4}$$

Consequentemente $\mathbf{Y} = (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$ com $\Lambda = (\lambda E p_1, \lambda E p_2, \lambda E p_3, \lambda E p_{13})$, com função de probabilidade conjunta deifinida por:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x_{13} \in D} \frac{e^{-E\lambda}(\lambda E p_1)^{y_1 - x_{13}}(\lambda E p_2)^{y_2}(\lambda E p_3)^{y_3 - x_{13}}(\lambda E p_{13})^{x_{13}}}{(y_1 - x_{13})!(y_2)!(y_3 - x_{13})!x_{13}!} \tag{4.5}$$

onde a soma é dada por $D \subset \mathcal{N}$ que é definido como:

$$D = [x_{13} \in \mathcal{N} : \{x_{13} \leq y_1\} \cap \{x_{13} \leq y_3\}].$$

O modelo 3 leva em consideração que só existem notificações em comum na base de dados dos boletins de ocorrência com a base de dados do SINASC, foi considerado que não existe probabilidade de ocorrer registro em comum entre as bases de boletins de ocorrência e SIM , ou seja, $p_{13} = 0$. Com isso, obtém-se o modelo 3:

$$\begin{aligned} Y_1 &= X_1 + X_{12}, \\ Y_2 &= X_2 + X_{12}, \\ Y_3 &= X_3. \end{aligned} \tag{4.6}$$

Consequentemente $\mathbf{Y} = (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$ com $\Lambda = (\lambda E p_1, \lambda E p_2, \lambda E p_3, \lambda E p_{12})$, com função de probabilidade conjunta deifinida por:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x_{12} \in T} \frac{e^{-\lambda E}(\lambda E p_1)^{y_1 - x_{12}}(\lambda E p_2)^{y_2 - x_{12}}(\lambda E p_3)^{y_3}(\lambda E p_{12})^{x_{12}}}{(y_1 - x_{12})!(y_2 - x_{12})!(y_3)!x_{12}!} \tag{4.7}$$

onde a soma é dada por $T \subset \mathcal{N}$ que é definido como:

$$T = [(x_{12} \in \mathcal{N} : \{x_{12} \leq y_1\} \cap \{x_{12} \leq y_2\}].$$

O modelo 4 é considerado o mais simples, uma vez que foi estipulado $p_{12} = p_{13} = 0$, ou seja, que não existe chance de ocorrer registros em comum nas bases analisadas. Esse modelo significa um grande número de subnotificações, dado que as vítimas não registraram a ocorrência na delegacia de polícia. Com isso, obtém-se o modelo 4:

$$\begin{aligned} Y_1 &= X_1, \\ Y_2 &= X_2, \\ Y_3 &= X_3. \end{aligned} \tag{4.8}$$

Consequentemente $\mathbf{Y} = (Y_1, Y_2, Y_3)' \sim 3 - Poisson(\Lambda)$ com $\Lambda = (\lambda E p_1, \lambda E p_2, \lambda E p_3,)$, com função de probabilidade conjunta deifinida por:

$$P(\mathbf{Y} = \mathbf{y}) = \frac{e^{-\lambda E}(\lambda E p_1)^{y_1}(\lambda E p_2)^{y_2}(\lambda E p_3)^{y_3}}{y_1! y_2! y_3!} \tag{4.9}$$

4.2 Função de verossimilhança

Seja $\mathbf{Y} = (\mathbf{Y_1}, \cdots, \mathbf{Y_n})$ uma amostra aleatória do quantitativo de estupro de vulnerável, em que n denota o número total de cidades analisadas. A partir das distribuições definidas em (4.2), (4.5), (4.7) e (4.9) é possível construir as funções de verossimilhança, dadas por:

- Modelo 1

$$\begin{aligned} L(\lambda, p_1, p_2, p_3, p_{12}, p_{13}) &= \prod_{i=1}^{n} \sum_{(x_{12}, x_{13}) \in C} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i} - x_{12i} - x_{13i}}(\lambda E_i p_2)^{y_{2i} - x_{12i}}}{(y_{1i} - x_{12i} - x_{13i})!(y_{2i} - x_{12i})!} \\ &\times \frac{(\lambda E_i p_3)^{y_{3i} - x_{13i}}(\lambda E_i p_{12})^{x_{12i}}(\lambda E_i p_{13})^{x_{13i}}}{(y_{3i} - x_{13i})! x_{12i}! x_{13i}!} \end{aligned} \tag{4.10}$$

- Modelo 2:

$$L(\lambda, p_1, p_2, p_3, p_{13}) = \prod_{i=1}^{n} \sum_{x_{13} \in D} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}-x_{13i}}(\lambda E_i p_2)^{y_{2i}}(\lambda E_i p_3)^{y_{3i}-x_{13i}}}{(y_{1i}-x_{13i})!(y_{2i})!(y_{3i}-x_{13i})!} \times \frac{(\lambda E_i p_{13})^{x_{13i}}}{x_{13i}!} \quad (4.11)$$

- Modelo 3:

$$L(\lambda, p_1, p_2, p_3, p_{12}) = \prod_{i=1}^{n} \sum_{x_{12} \in T} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}-x_{12i}}(\lambda E_i p_2)^{y_{2i}-x_{12i}}(\lambda E_i p_3)^{y_{3i}}}{(y_{1i}-x_{12i})!(y_{2i}-x_{12i})!(y_{3i})!} \times \frac{(\lambda E_i p_{12})^{x_{12i}}}{x_{12i}!} \quad (4.12)$$

- Modelo 4

$$L(\lambda, p_1, p_2, p_3) = \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}}(\lambda E_i p_2)^{y_{2i}}(\lambda E_i p_3)^{y_{3i}}}{y_{1i}!y_{2i}!y_{3i}!} \quad (4.13)$$

4.3 Procedimentos de inferência

A abordagem escolhida para fazer a inferência dos parâmetros do modelo neste trabalho é a bayesiana. Esta abordagem ocorre na atualização do conhecimento a priori com a informação vinda das observações, expresso pela distribuição a posteriori. Mesmo que sua solução seja analiticamente inviável, é possível aproximar a distribuição a posteriori recorrendo a métodos de simulação. Em particular, para facilitar o desempenho computacional do modelo abordado nesta dissertação foi utilizado o artifício de aumento de dados definido por [19].

4.3.1 Função de verossimilhança aumentada

Introduzido por [19], a ideia do aumento de dados, consiste em se adotar variáveis latentes na função de verossimilhança de forma que a estrutura probabilística do processo se mantenha coerente com a proposta original. Para os modelos (4.1), (4.4) e (4.6), foi assumido X_{12} e X_{13} como variáveis latentes. Desta forma, as funções de verossimilhança aumentada são dadas por:

- Modelo 1

$$L(\lambda, p_1, p_2, p_3, p_{12}, p_{13}, \mathbf{x_{12}}, \mathbf{x_{13}}) = \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}-x_{12i}-x_{13i}}(\lambda E_i p_2)^{y_{2i}-x_{12i}}}{(y_{1i}-x_{12i}-x_{13i})!(y_{2i}-x_{12i})!} \times \frac{(\lambda E_i p_3)^{y_{3i}-x_{13i}}}{(y_{3i}-x_{13i})!} \quad (4.14)$$

- Modelo 2

$$L(\lambda, p_1, p_2, p_3, p_{13}, \mathbf{x_{13}}) = \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}-x_{13i}}(\lambda E_i p_2)^{y_{2i}}}{(y_{1i}-x_{13i})!y_{2i}!} \times \frac{(\lambda E_i p_3)^{y_{3i}-x_{13i}}}{(y_{3i}-x_{13i})!} \quad (4.15)$$

- Modelo 3

$$L(\lambda, p_1, p_2, p_3, p_{12}, \mathbf{x_{12}}) = \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i}-x_{12i}}(\lambda E_i p_3)^{y_{3i}}}{(y_{1i}-x_{12i})!y_{3i}!} \times \frac{(\lambda E_i p_2)^{y_{2i}-x_{12i}}}{(y_{2i}-x_{12i})!} \quad (4.16)$$

Perceba que a introdução de variáveis latentes tem como principal vantagem o fato de não precisar recorrer aos complicados somatórios dados nas funções (4.10), (4.11) e (4.12).

4.4 Formulação da distribuição a posteriori

Para realizar inferência bayesiana é necessário encontrar a distribuição a posteriori do vetor de parâmetros $\psi = (\lambda, p_1, p_2, p_3, p_{12}, p_{13})$.

Como $(p_1, p_2, p_3, p_{12}, p_{13})$ são parâmetros que indicam probabilidades e λ indica uma taxa maior do que zero, as seguintes distribuições são coerentes aos referidos parâmetros:

$$\lambda \sim Gama(a, b) \quad (4.17)$$

$$(p_1, p_2, p_3, p_{12}, p_{13}) \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3, \alpha_{12}, \alpha_{13}) \quad (4.18)$$

em que $\alpha_1, \alpha_2, \alpha_3, \alpha_{12}, \alpha_{13}, a$ e b são hiperparâmetros positivos conhecidos.

Em função da independência dos parâmetros, as prioris acima podem ser combinadas numa distribuição a priori conjunta $\pi(\psi)$:

$$\pi(\psi) \propto \lambda^{a-1} \exp\{-\lambda b\} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_{12}^{\alpha_{12} - 1} p_{13}^{\alpha_{13} - 1} \tag{4.19}$$

Ao se associar a referida priori com as funções de verossimilhança identificada em (4.14), (4.15), (4.16) e (4.13), teremos as seguintes posterioris:

- Modelo 1

$$\begin{aligned} f(\psi, \mathbf{X_{12}}, \mathbf{X_{13}} \mid \mathbf{Y}) &\propto \pi(\psi) L(\lambda, p_1, p_2, p_3, p_{12}, p_{13}, \mathbf{x_{12}}, \mathbf{x_{13}}) f(\mathbf{X_{12}} \mid \psi) \\ &\times f(X_{13} \mid \psi) \\ &\propto \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i} - x_{12i} - x_{13i}} (\lambda E_i p_2)^{y_{2i} - x_{12i}} (\lambda E_i p_3)^{y_{3i} - x_{13i}}}{(y_{1i} - x_{12i} - x_{13i})!(y_{2i} - x_{12i})!(y_{3i} - x_{13i})!} \\ &\times \frac{(\lambda E_i p_{12})^{x_{12i}} (\lambda E_i p_{13})^{x_{13i}}}{x_{12i}! x_{13i}!} \\ &\times \lambda^{a-1} \exp\{-\lambda b\} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_{12}^{\alpha_{12} - 1} p_{13}^{\alpha_{13} - 1} \end{aligned} \tag{4.20}$$

- Modelo 2

$$\begin{aligned} f(\psi, \mathbf{X_{13}} \mid \mathbf{Y}) &\propto \pi(\psi) L(\lambda, p_1, p_2, p_3, p_{13}, \mathbf{x_{13}}) f(\mathbf{X_{13}} \mid \psi) \\ &\propto \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i} - x_{13i}} (\lambda E_i p_2)^{y_{2i}} (\lambda E_i p_3)^{y_{3i} - x_{13i}} (\lambda E_i p_{13})^{x_{13i}}}{(y_{1i} - x_{13i})!(y_{3i} - x_{13i})! x_{13i}!} \\ &\times \lambda^{a-1} \exp\{-\lambda b\} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_{13}^{\alpha_{13} - 1} \end{aligned} \tag{4.21}$$

- Modelo 3

$$\begin{aligned} f(\psi, \mathbf{X_{12}} \mid \mathbf{Y}) &\propto \pi(\psi) L(\lambda, p_1, p_2, p_3, p_{12}, \mathbf{x_{12}}) f(\mathbf{X_{12}} \mid \psi) \\ &\propto \prod_{i=1}^{n} \frac{exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1i} - x_{12i}} (\lambda E_i p_2)^{y_{2i} - x_{12i}} (\lambda E_i p_3)^{y_3 i} (\lambda E_i p_{12})^{x_{12i}}}{(y_{1i} - x_{12i})!(y_{2i} - x_{12i})! x_{12i}!} \\ &\times \lambda^{a-1} \exp\{-\lambda b\} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_{12}^{\alpha_{12} - 1} \end{aligned} \tag{4.22}$$

- Modelo 4

$$
\begin{aligned}
f(\psi \mid \mathbf{Y}) &\propto \pi(\psi) L(\lambda, p_1, p_2, p_3) \\
&\propto \prod_{i=1}^{n} exp(-\lambda E_i)(\lambda E_i p_1)^{y_{1_i}} (\lambda E_i p_2)^{y_{2_i}} (\lambda E p_3)^{y_{3_i}} \\
&\times \lambda^{a-1} \exp\{-\lambda b\} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1}
\end{aligned} \tag{4.23}
$$

E as distribuições condicionais completas são dadas por:

- Modelo 1
  - Distribuição condicional completa para λ

$$
\lambda \sim Gama\left(\sum_{i=1}^{n}(x_{1i} + x_{2i} + x_{3i} - x_{12i} - x_{13i}) + a, \sum_{i=1}^{n} E_i + b\right) \tag{4.24}
$$

  - Distribuição condicional completa para $(p_1, p_2, p_3, p_{12}, p_{13})$

$$
(p_1, p_2, p_3, p_{12}, p_{13}) \sim Dirichlet\,(\alpha_1', \alpha_2', \alpha_3', \alpha_{12}', \alpha_{13}') \tag{4.25}
$$

    onde

$$
\alpha_1' = \sum_{i=1}^{n}(y_{1i} - x_{12i} - x_{13i}) + \alpha_1 \tag{4.26}
$$

$$
\alpha_2' = \sum_{i=1}^{n}(y_{2i} - x_{12i}) + \alpha_2 \tag{4.27}
$$

$$
\alpha_3' = \sum_{i=1}^{n}(y_{3i} - x_{13i}) + \alpha_3 \tag{4.28}
$$

$$
\alpha_{12}' = \sum_{i=1}^{n} x_{12i} + \alpha_{12} \tag{4.29}
$$

$$
\alpha_{13}' = \sum_{i=1}^{n} x_{13i} + \alpha_{13} \tag{4.30}
$$

  - Distribuição condicional completa para x_{12}

$$
f(x_{12i}) \propto \left[\frac{p_{12}}{\lambda E_i p_1 p_2}\right]^{x_{12i}} \frac{1}{(y_{1i} - x_{12i} - x_{13i})!(y_{2i} - x_{12i})! x_{12}!} \tag{4.31}
$$

– Distribuição condicional completa para x_{13}

$$f(x_{13i}) \propto \left[\frac{p_{13}}{\lambda E_i p_1 p_3}\right]^{x_{13i}} \frac{1}{(y_{1i} - x_{12i} - x_{13i})!(y_{3i} - x_{13i})! x_{13}!} \tag{4.32}$$

- Modelo 2

  – Distribuição condicional completa para λ

$$\lambda \sim Gama\left(\sum_{i=1}^{n}(x_{1i} + x_{2i} + x_{3i} - x_{13i}) + a, \sum_{i=1}^{n} E_i + b\right) \tag{4.33}$$

  – Distribuição condicional completa para (p_1, p_2, p_3, p_{13})

$$(p_1, p_2, p_3, p_{13}) \sim Dirichlet\,(\alpha'_1, \alpha'_2, \alpha'_3, \alpha'_{13}) \tag{4.34}$$

  onde

$$\alpha'_1 = \sum_{i=1}^{n}(y_{1i} - x_{13i}) + \alpha_1 \tag{4.35}$$

$$\alpha'_2 = \sum_{i=1}^{n} y_{2i} + \alpha_2 \tag{4.36}$$

$$\alpha'_3 = \sum_{i=1}^{n}(y_{3i} - x_{13i}) + \alpha_3 \tag{4.37}$$

$$\alpha'_{13} = \sum_{i=1}^{n} x_{13i} + \alpha_{13} \tag{4.38}$$

  – Distribuição condicional completa para x_{13}

$$f(x_{13i}) \propto \left[\frac{p_{13}}{\lambda E_i p_1 p_3}\right]^{x_{13i}} \frac{1}{(y_{1i} - x_{13i})!(y_{3i} - x_{13i})! x_{13}!} \tag{4.39}$$

- Modelo 3

  – Distribuição condicional completa para λ

$$\lambda \sim Gama\left(\sum_{i=1}^{n}(x_{1i} + x_{2i} + x_{3i} - x_{12i}) + a, \sum_{i=1}^{n} E_i + b\right) \tag{4.40}$$

  – Distribuição condicional completa para (p_1, p_2, p_3, p_{12})

$$(p_1, p_2, p_3, p_{12}) \sim Dirichlet\,(\alpha'_1, \alpha'_2, \alpha'_3, \alpha'_{12}) \tag{4.41}$$

onde

$$\alpha'_1 = \sum_{i=1}^{n}(y_{1i} - x_{12i}) + \alpha_1 \tag{4.42}$$

$$\alpha'_2 = \sum_{i=1}^{n}(y_{2i} - x_{12i}) + \alpha_2 \tag{4.43}$$

$$\alpha'_3 = \sum_{i=1}^{n} y_{3i} + \alpha_3 \tag{4.44}$$

$$\alpha'_{12} = \sum_{i=1}^{n} x_{12i} + \alpha_{12} \tag{4.45}$$

$$\tag{4.46}$$

– Distribuição condicional completa para x_{12}

$$f(x_{12i}) \propto \left[\frac{p_{12}}{\lambda E_i p_1 p_2}\right]^{x_{12i}} \frac{1}{(y_{1i} - x_{12i})!(y_{2i} - x_{12i})!x_{12}!} \tag{4.47}$$

- Modelo 4

– Distribuição condicional completa para λ

$$\lambda \sim Gama\left(\sum_{i=1}^{n}(x_{1i} + x_{2i} + x_{3i}) + a, \sum_{i=1}^{n} E_i + b\right) \tag{4.48}$$

– Distribuição condicional completa para (p_1, p_2, p_3)

$$(p_1, p_2, p_3) \sim Dirichlet\left(\sum_{i=1}^{n} y_{1i}\alpha_1, \sum_{i=1}^{n} y_{2i} + \alpha_2, \sum_{i=1}^{n} y_{3i} + \alpha_3\right) \tag{4.49}$$

Note que, as distribuições a posteriori encontradas em (4.20), (4.21), (4.22) e (4.23) são discretas com suporte finito. Portanto, as mesmas podem trivialmente ser simuladas a partir de uma distribuição discreta [9].

Capítulo 5

Aplicação

Neste capítulo, foram aplicados os modelos propostos aos dados no Capítulo 3 para avaliar o desempenho da metodologia proposta. Em particular, o objetivo principal é verificar estimativas dos parâmetros e a capacidade dos modelos propostos no Capítulo 4 se ajustarem melhor aos conjuntos de dados em análise.

5.1 Resultados

Os sumários dos resultados das distribuições a posteriori serão apresentados nas próximas seções, de maneira que foi aplicado os modelos propostos no Capítulo 4 para os dados de registros de notificações apresentados nas Tabelas 3.1, 3.2 e 3.3. Assumindo distribuições a priori pouco informativas, ou seja, uma priori com muita variabilidade para λ e $(p_1, p_2, p_3, p_{12}, p_{13})$ com hiperparâmetros cujos valores foram $a = 0.1, b = 0.1, \alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1, \alpha_{12} = 1, \alpha_{13} = 1$, simulamos no `R` [16] amostras de Gibbs, nas quais foram geradas quatro cadeias de 5.000 iterações para os parâmetros, as primeiras 2.500 iterações foram descartadas (*burn-in*). Nesse contexto, foi utilizada a técnica proposta por [7], para monitorar a convergência das cadeias, através do pacote Coda [15], disponível no `R`. Foi considerado que a cadeia convergiu quando $\hat{R} < 1.1$ para todos os parâmetros.

Os modelos propostos foram aplicados aos dados de notificações entre as vítimas com idade inferior a 14, 13 e 12 anos. A Figura 5.1 apresenta o comportamento das cadeias geradas para os parâmetros do modelo 1. Através dessa figura, percebe-se que há uma certa uniformidade nos traços a posteriori dos parâmetros estimados, indicando possível convergência. Com a aplicação do método proposto por [7] , foi verificado que as cadeias geradas apresentaram convergência, pois os valores de $\hat{R}$ foram inferiores a 1.1.

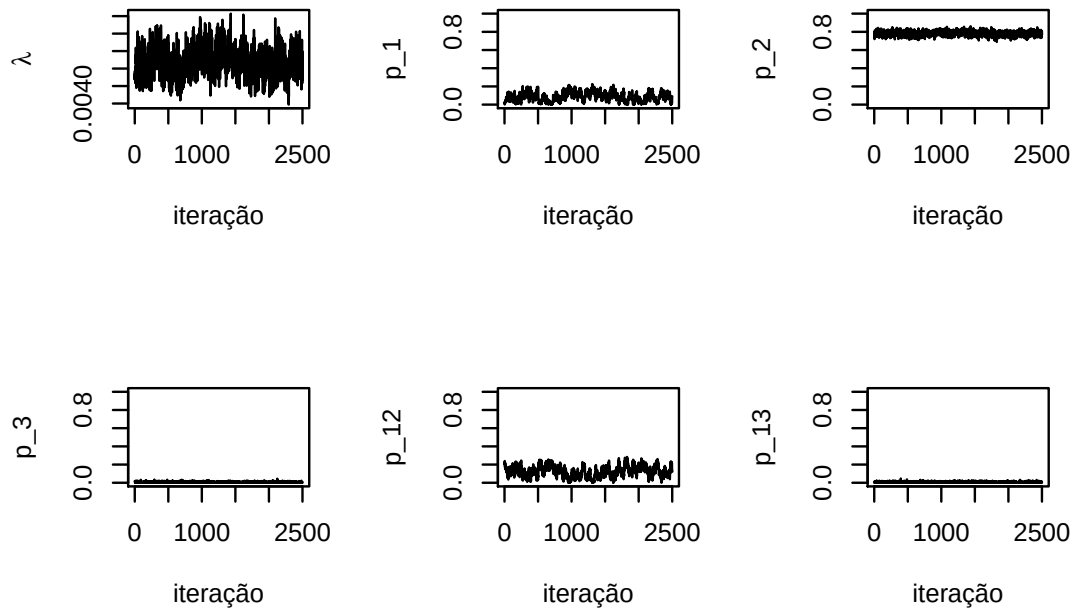


Figura 5.1: Traços a posteriori para os parâmetros do modelo 1

A Tabela 5.1 mostra o DIC dos 4 modelos para cada faixa etária.

Tabela 5.1: Seleção de modelos com base no critério de informação Bayesianos DIC

	DIC		
	Menores de 14 anos	Menores de 13 anos	Menores de 12 anos
Modelo 1	121.1217	95.21172	76.41782
Modelo 2	122.944	94.98192	74.1626
Modelo 3	125.011	97.92479	77.15151
Modelo 4	146.5348	113.2444	82.38082

Tabela 5.2: Resumo da distribuição posteriori. Média e intervalo de credibilidade 95% (IC)

	Parâmetros	Média	IC
Menores de 14 anos	λ	51.84	[44.56 ; 61.04]
	p_1	0.084389429	[0.0086426 ; 0.167253]
	p_2	0.787061201	[0.7380534 ; 0.831299]
	p_3	0.006570230	[0.0001812 ; 0.021223]
	p_{12}	0.115103049	[0.0301478 ; 0.208993]
	p_{13}	0.00687609	[0.0003624 ; 0.020284]
Menores de 13 anos	λ	19.82	[16.24 ; 23.69]
	p_1	0.393557772	[0.3104636 ; 0.507448]
	p_2	0.571762698	[0.4625337 ; 0.653253]
	p_3	0.017491231	[0.0005893 ; 0.045480]
	p_{13}	0.017188299	[0.0002528 ; 0.044141]
Menores de 12 anos	λ	7.4	[5.35 ; 10.21]
	p_1	0.5827756713	[0.4549157 ; 0.729493]
	p_2	0.3564815293	[0.2437254 ; 0.477462]
	p_3	0.0294572863	[0.0011741 ; 0.102387]
	p_{13}	0.0312855131	[0.0017514 ; 0.108326]

Na Tabela 5.1, encontra-se o valor do DIC para os modelos. Comparando-se os valores, pode-se dizer que o modelo 1 se ajustou melhor aos dados da vítimas com idade inferior a 14 anos, o modelo 3 se ajustou melhor aos dados da vítimas com idade inferior a 13 anos e o modelo 2 se ajustou melhor aos dados da vítimas com idade inferior a 12 anos.

Com base no resultado apresentados na tabela 5.2, é possível concluir que as taxas médias de estupro de vulnerável para cada 10.000 habitantes para vítimas abaixo de 14, 13 e 12 anos são de 51.84, 19.82 e 7.4 respectivamente e as estimativas das probabilidades médias do crime de estupro de vulnerável ser subnotificada são de 79%, 59% e 39% para vítimas abaixo de 14, 13 e 12 anos respectivamente.

5.2 Análise preditiva a posteriori

A habilidade de predição e bondade de ajuste são características distintas do modelo. O modelo pode explicar e predizer adequadamente as observações usadas na sua construção, entretanto, pode fazer predições ruins para observações futuras ou fora do intervalo da variável resposta observada [17]. As Figuras 5.2 a 5.4 apresentam a predição dos modelos selecionados na Seção 5.1 para cada base de dados analisada.

Os intervalos de credibilidade de 95% apresentados foram construídos a partir da simulação de 2.334 amostras da distribuição preditiva a posteriori.

Pelas Figuras de 5.2 a 5.4 observa-se que as estimativas marginais obtidas pelo MCMC apresentaram boa aderência, tem-se que em grande parte o verdadeiro valor de $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ se encontra dentro do intervalo de credibilidade, note que apenas na cidade de Atalaia do Norte para as vítimas abaixo de 14 anos na base de dados do SINASC o valor verdadeiro não se encontra dentro do intervalo. Isso revela que essa abordagem se mostra, a princípio, adequada na análise bayesiana.

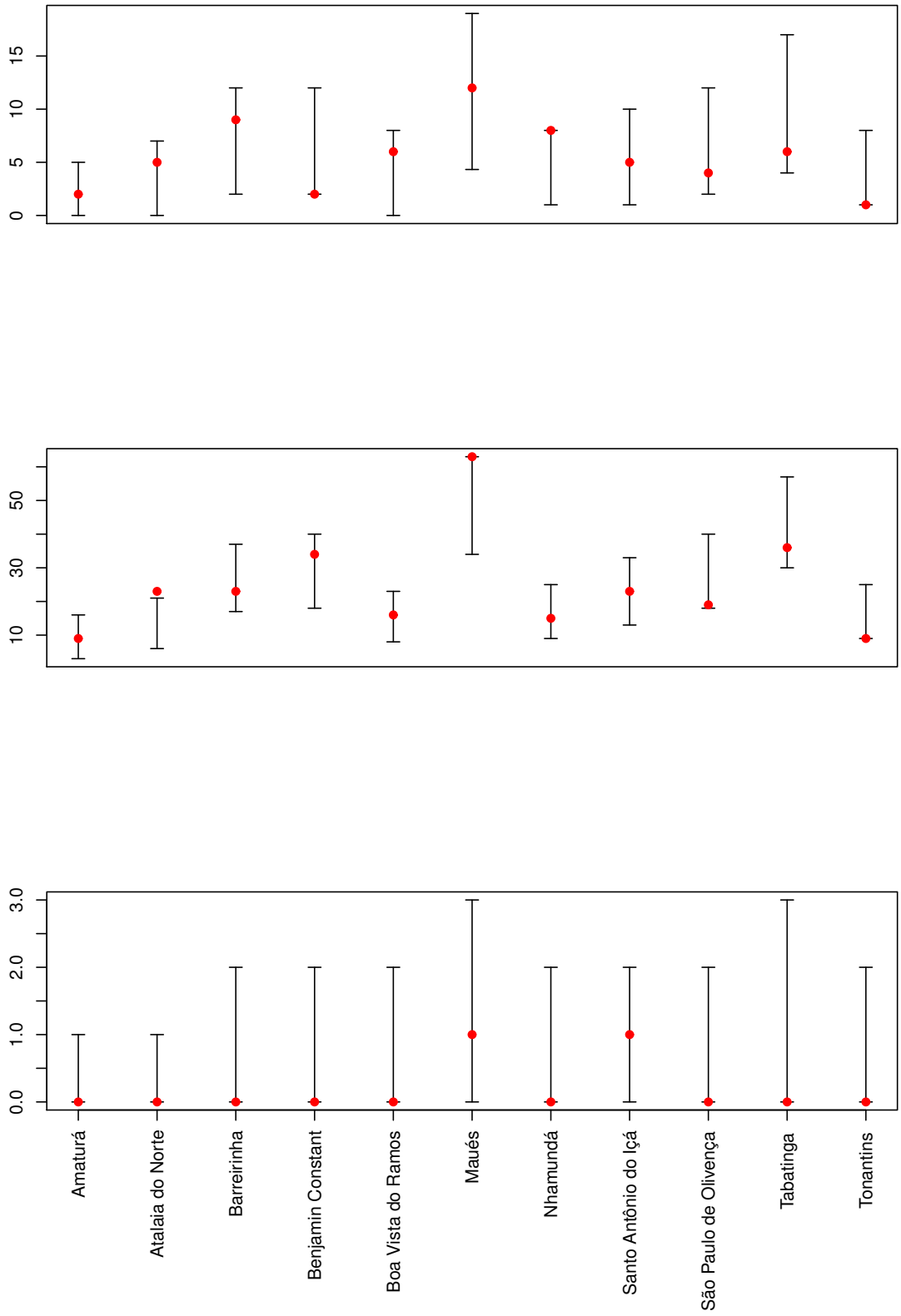


Figura 5.2: De cima para baixo: Sumário dos dados preditivos de vítimas abaixo de 14 anos para delegacia, SINASC e SIM. O círculo vermelho representa o valor verdadeiro do número de vítimas e a linha é o intervalo de predição de 95%.

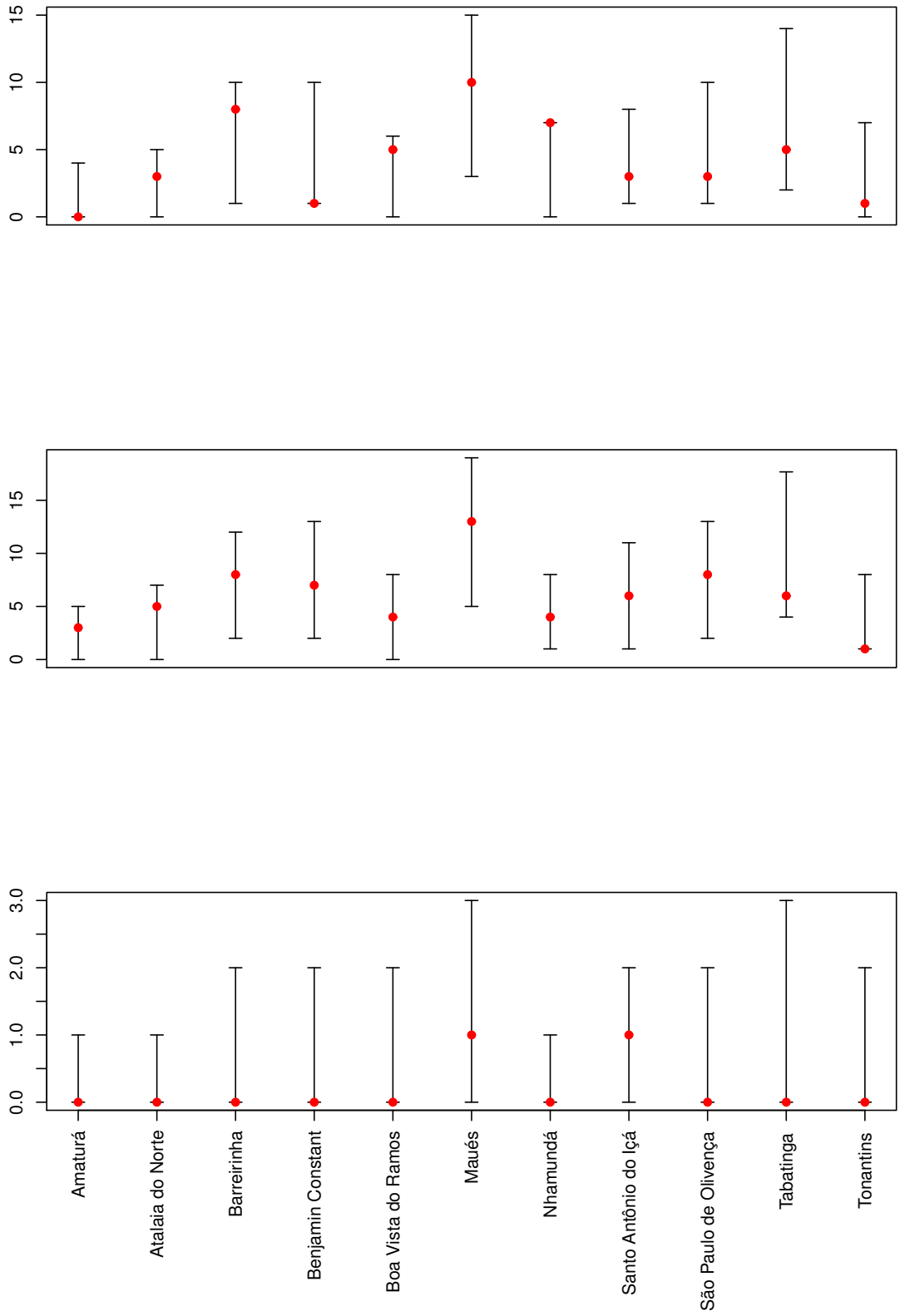


Figura 5.3: De cima para baixo: Sumário dos dados preditivos de vítimas abaixo de 13 anos para delegacia, SINASC e SIM. O círculo vermelho representa o valor verdadeiro do número de vítimas e a linha é o intervalo de predição de 95%.

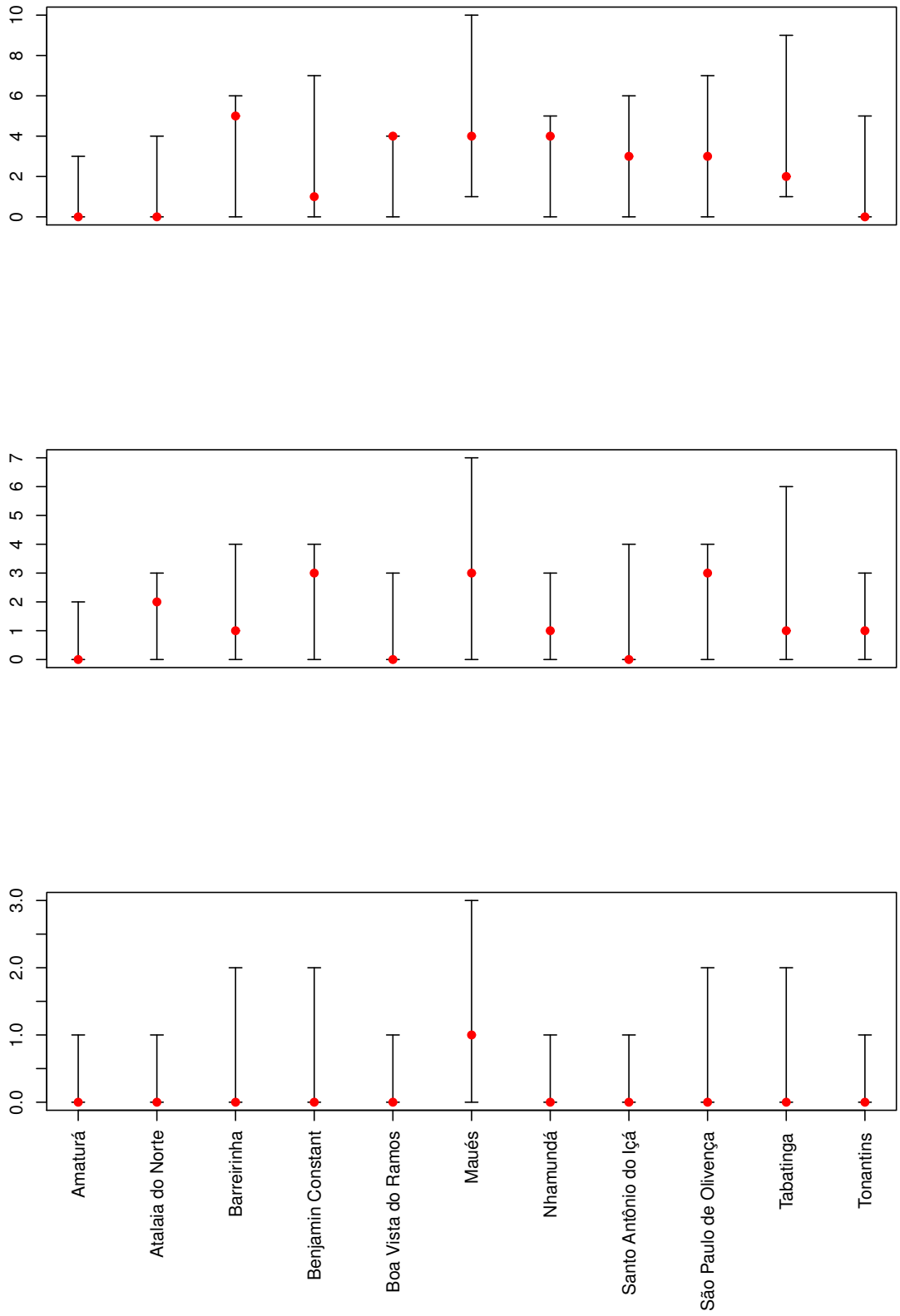


Figura 5.4: De cima para baixo: Sumário dos dados preditivos de vítimas abaixo de 12 anos para delegacia, SINASC e SIM. O círculo vermelho representa o valor verdadeiro do número de vítimas e a linha é o intervalo de predição de 95%.

Capítulo 6

Conclusão

A metodologia bayesiana, diante dos resultados obtidos, mostrou-se eficiente para estimar da taxa de estupro de vulnerável nas cidades analisadas por meio do modelo $3 - Poisson$. Este modelo é uma ferramenta importante para corrigir as taxas de estupro de vulnerável, uma vez que as mesmas dificilmente podem ser calculadas de modo adequado utilizando a identificação da vítima entre os bancos.

Realizando um panorama geral, pode-se afirmar que as subnotificações diminuíram conforme o corte de idade foi decrescendo. O modelo 1 apresentou o melhor ajuste aos dados juntamente com uma boa habilidade de predição com os dados das vítimas menores de 14 anos , ou seja, foi o único grupo que apresentou probabilidade de ocorrer registros em comum com as duas bases do SUS. Este modelo evidenciou uma estimativa intervalar para a taxa de estupro de vulnerável para cada 10 mil habitantes de [44.56 ; 61.04] para as cidades analisadas, enquanto que [3] apontou um taxa de estupro de 23,2% para cada 100 mil habitantes. Nessa perspectiva,[2] estima que sejam efetivamente notificados apenas 10% dos crimes de estupro , este trabalho mostrou que a notificação é variável por idade, sendo estimada entre 21%, na faixa de 14 anos, até 61% na faixa dos 12 anos.

Como trabalho futuro propõe-se acrescentar a base de dados do Sistema de Informação de Agravos de Notificação (SINAN) como uma variável do modelo de Poisson, para estimar a taxa de estupros de vulnerável levando em consideração as vítmias que não passaram por um processo de gravidez.

Referências Bibliográficas

[1] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[2] D Cerqueira and DSC Coelho. Estupro no brasil: uma radiografia segundo os dados da saúde. *Brasília: IPEA*, 2014.

[3] Fórum Brasileiro de Segurança Pública. Anuário brasileiro de segurança pública. 2017.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[5] D. Gamerman and H.F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2006. ISBN 9781584885870.

[6] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[7] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[8] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[9] James E Gentle. *Computational statistics*. Springer Science & Business Media, 2009.

[10] John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991.

[11] Kazutomo Kawamura. The structure of trivariate poisson distribution. In *Kodai Mathematical Seminar Reports*, volume 28, pages 1–8. Department of Mathematics, Tokyo Institute of Technology, 1976.

[12] Kazutomo Kawamura. The structure of multivariate poisson distribution. *Kodai Mathematical Journal*, 2(3):337–345, 1979.

[13] S Loukas and CD Kemp. On computer sampling from trivariate and multivariate discrete distributions: Multivariate discrete distributions. *Journal of Statistical Computation and Simulation*, 17(2):113–123, 1983.

[14] H.S. Migon, D. Gamerman, and F. Louzada. *Statistical Inference: An Integrated Approach, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2014. ISBN 9781439878804.

[15] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006. URL `http://CRAN.R-project.org/doc/Rnews/`.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

[17] Daniel Sorensen and Daniel Gianola. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer Science & Business Media, 2007.

[18] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639, 2002.

[19] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

[20] Efthymios G Tsionas. Bayesian analysis of the multivariate poisson distribution. *Communications in Statistics-Theory and Methods*, 28(2):431–451, 1999.

[21] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 1983.