UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos bayesianos zero-modificados para séries temporais de contagem

Caroline Mendes de Assis

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Caroline Mendes de Assis

Modelos bayesianos zero-modificados para séries temporais de contagem

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Marinho Gomes de Andrade Filho

USP – São Carlos
Maio de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M538m

Mendes de Assis, Caroline
Modelos bayesianos zero-modificados para séries temporais de contagem / Caroline Mendes de Assis; orientador Marinho Gomes de Andrade Filho. -- São Carlos, 2020.
111 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Modelos zero-modificados. 2. Dados de contagem. 3. Distribuição COM-Poisson. 4. Modelos generalizados ARMA.. I. Gomes de Andrade Filho, Marinho, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Caroline Mendes de Assis

Bayesian zero-modified models for count time series

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Marinho Gomes de Andrade Filho

USP – São Carlos
May 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Caroline Mendes de Assis, realizada em 30/04/2020:

Prof. Dr. Marinho Gomes de Andrade Filho
USP

Profa. Dra. Juliana Cobre
ICMC/USP

Profa. Dra. Thelma Sáfadi
UFLA

Profa. Dra. Sandra Cristina de Oliveira
UNESP

Prof. Dr. Eder Angelo Milani
UFG

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Marinho Gomes de Andrade Filho Juliana Cobre, Thelma Sáfadi, Sandra Cristina de Oliveira, Eder Angelo Milani e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Marinho Gomes de Andrade Filho

Dedico este trabalho ao meu filho Bernardo.

AGRADECIMENTOS

Gostaria de agradecer à minha família, especialmente a meus pais Ana e Francisco, irmãos Catherine e Bruno e a meu tio Estêvão. À memória de minha avó Fátima, que tanto me apoiou durante a trajetória do doutorado. Agradeço ainda ao meu esposo Raul, que ao longo da escrita desta tese passou de namorado a esposo e agora pai de nosso filho.

Ao professor e orientador Marinho Gomes de Andrade Filho, pela orientação, conversas esclarecedoras e confiança depositada em mim para a realização deste trabalho.

Aos professores do PIPGEs, que contribuíram para minha formação acadêmica. Aos membros da comissão examinadora, que enriqueceram este trabalho com sugestões e comentários.

Aos meus amigos e colegas de turma pelos momentos de estudo e distração. Aos amigos da Assessoria de Gestão Estratégica do Superior Tribunal Militar: Raissa, Ingrid, Mosair, André, Carolina, Rafael, Estanislau e Rodrigo, que muito me apoiaram para que eu obtivesse essa conquista.

A todos aqueles que contribuíram direta ou indiretamente para a produção deste trabalho, agradeço imensamente.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“De tudo, ficaram três coisas:
a certeza de que ele estava sempre começando,
a certeza de que era preciso continuar
e a certeza de que seria interrompido antes de terminar.
Fazer da interrupção um caminho novo.
Fazer da queda um passo de dança,
do medo uma escada, do sono uma ponte,
da procura um encontro.”
(Fernando Sabino)

RESUMO

ASSIS, C. M. **Modelos bayesianos zero-modificados para séries temporais de contagem**. 2020. 111 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Neste trabalho são apresentados dois modelos bayesianos zero-modificados (ZM) para séries temporais de contagem: Poisson ARMA zero-modificado e COM-Poisson ARMA zero-modificado. O segundo modelo permite uma flexibilidade maior por possuir um parâmetro adicional que comporta dados com maior sobredispersão ou subdispersão em relação ao modelo Poisson ARMA ZM. Os modelos são ilustrados por meio de aplicação em dados artificiais e em dois conjuntos de dados reais. Tanto o modelo Poisson ARMA ZM quanto o modelo COM-Poisson ARMA ZM se mostraram competitivos para modelar dados de contagem zero-modificados, tendo sido estudado o ajuste dos modelos aos dados por meio da análise preditiva a posteriori. A comparação de modelos foi realizada por meio do critério de informação da deviância (DIC). Finalmente, foi realizado um estudo de previsão para seis períodos à frente. De maneira geral, o modelo COM-Poisson ARMA ZM, apesar de possuir um parâmetro adicional em relação ao modelo Poisson ARMA ZM, obteve valores de DIC próximos aos do modelo Poisson ARMA ZM. Como o modelo COM-Poisson ARMA ZM possui como caso particular o modelo Poisson ARMA ZM, tendo a vantagem de ser mais flexível, o modelo COM-Poisson ARMA ZM é proposto como uma alternativa para dados de contagem com modificação na contagem de zeros.

Palavras-chave: Modelos zero-modificados. Dados de contagem. Distribuição COM-Poisson. Modelos generalizados ARMA.

ABSTRACT

ASSIS, C. M. **Bayesian zero-modified models for count time series**. 2020. 111 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

This work presents two Bayesian zero-modified (ZM) models for count time series: zero-modified Poisson ARMA and zero-modified COM-Poisson ARMA. The latter allows a greater flexibility since it has an aditional parameter which accomodates greater subdispersion or overdispersion in comparison with the ZM Poisson ARMA model. The models are applied to simulated data and two real data sets. Both ZM Poisson ARMA and ZM COM-Poisson ARMA performed very well in zero-modified data. The goodness of fit was studied using posterior predictive checks. Model comparison was done using the deviance information criterion (DIC). Finally, a forecast study of six-steps-ahead was performed. In general, the ZM COM-Poisson model, although having an aditional parameter in comparison with the ZM Poisson ARMA model, showed DIC values similar to the DIC values of the ZM Poisson ARMA model. Since the ZM COM-Poisson ARMA model has the ZM Poisson ARMA model as a particular case, having the advantage of being more flexible, the ZM COM-Poisson ARMA model is proposed as an alternative to zero-modified count data.

Keywords: Zero-modified models. Count data. COM-Poisson distribution. Generalized ARMA models..

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Motivação

Séries temporais de contagem podem ser encontradas em diversas situações e áreas de estudo. Em epidemiologia pode-se estudar notificações mensais de poliomielite (ZEGER, 1988); na análise e prevenção de acidentes pode-se analisar a contagem anual de fatalidades em rodovias (QUDDUS, 2008); na área de negócios de transportes é de interesse verificar o número de reservas por hora em um sistema de carros compartilhados ao longo de alguns meses (MÜLLER; BOGENBERGER, 2015).

Uma característica importante que pode ser encontrada nas séries temporais de contagem é a inflação ou deflação de observações iguais a zero. Por exemplo, registros semanais de casos de sífilis registrados em Maryland, EUA, no período de 2007 a 2010 contêm um número alto de observações iguais a zero (YANG, 2012). Como notado por Yang (2012), o excesso de observações iguais a zero é comum quando as taxas de infecção são baixas em contagens de episódios de doenças. Outro dado importante em relação a observações epidemiológicas é que elas são coletadas ao longo do tempo, e assim é importante considerar a correlação dos valores atuais com os valores passados ao estudar esses fenômenos.

Uma das distribuições mais conhecidas e utilizadas para modelar dados de contagem é a distribuição de Poisson. No entanto, tal distribuição pode não ser recomendada caso os dados possuam sub ou sobredispersão: casos em que a variância é menor ou maior que a média, respectivamente. Há casos também em que há um excesso de observações iguais a zero, mais do que seria explicado pela distribuição de Poisson.

1.2 Justificativa

A partir do que foi disposto na seção de motivação, a justificativa para este trabalho é a necessidade de modelos mais condizentes com dados que ocorrem na natureza, isto é, com deflação ou inflação de zeros, e ainda com a possibilidade de haver sub ou sobredispersão, já que em muitos casos o valor esperado é diferente da variância dos dados.

1.3 Objetivos

O objetivo geral deste trabalho é propor modelos bayesianos zero-modificados para séries temporais de contagem. Como objetivos específicos têm-se: apresentar uma abordagem bayesiana usando prioris não informativas para dois modelos série de potência ARMA ZM: Poisson ARMA ZM e COM-Poisson ARMA ZM; fazer uma comparação entre os dois modelos considerados por meio de dados artificiais e aplicações a dados reais e mostrar o procedimento para previsão em modelos da família série de potência ARMA ZM.

1.4 Revisão da literatura

Modelos de regressão que incorporam inflação de zeros foram estudados por Lambert (1992), Broek (1995), Hall (2000), Dietz e Böhning (2000), Lee, Wang e Yau (2001), assim como em Rodrigues (2003).

Lambert (1992) trata da distribuição proveniente da mistura entre uma variável degenerada no zero e uma distribuição de Poisson usual, que resulta nos modelos de Poisson zero inflacionados, ou ZIP. Sua aplicação se deu no contexto de contagem de falhas em placas de circuito impresso.

Broek (1995) utilizou um teste para averiguar se o número de zeros está inflacionado em relação ao número correspondente da distribuição de Poisson, e mostrou a utilidade do modelo ZIP ao aplicá-lo em dados de contagem de episódios de infecções no trato urinário de pacientes com HIV, em que muitos deles não haviam apresentado infecção urinária.

Dietz e Böhning (2000) avaliaram a prevenção da saúde bucal em um estudo mais geral, pois tratou do modelo ZMP, i.e., do modelo Poisson zero-modificado, havendo assim a possibilidade de incorporar dados com um número menor de zeros do que o esperado para uma distribuição de Poisson.

Rodrigues (2003) estudou distribuições zero inflacionadas em um contexto bayesiano, aplicando o modelo ZIP em um estudo de plantações de maçãs em meios diferentes de fotoperíodo e concentrações de um hormônio vegetal.

Conceição, Andrade e Louzada (2014) fizeram uma análise do modelo ZMP no contexto bayesiano utilizando medidas de divergência a posteriori. A aplicação foi feita em dados de incidência de terrorismo internacional.

Como mostrado, muitos estudos foram realizados no contexto de modelos de regressão. Quando nos voltamos para estudos de dados de contagem zero-modificados em séries temporais, o material bibliográfico torna-se mais escasso.

Alguns modelos de contagem para séries temporais se tornaram possíveis com os modelos autorregressivos e de médias móveis generalizados (GARMA), propostos por Benjamin, Rigby e Stasinopoulos (2003). A classe de modelos GARMA permite uma flexibilidade maior ao tratar de dados de séries temporais, pois a distribuição condicional dos dados pertence à família exponencial; não é restrita à distribuição normal, como no caso dos modelos autorregressivos e de médias móveis gaussianos (ARMA). Na especificidade de dados de contagem, as distribuições que podem ser usadas nos modelos GARMA são, por exemplo, Poisson e binomial negativa.

Uma compilação de modelos e técnicas para o estudo de séries temporais de contagem pode ser encontrada em Davis et al. (2015). Yau, Lee e Carrivick (2004) propuseram o modelo ZIP com estrutura autorregressiva de primeira ordem e aplicaram em dados de acidentes de trabalho observados de julho de 1988 a outubro de 1995.

Fazendo a conjunção entre os modelos de séries temporais para dados de contagem e permitindo a modificação no número de observações iguais a zero, há o trabalho de Yang (2012), que estudou os modelos dinâmicos ZIP e ZINB, do inglês binomial negativo inflacionado de zeros, em uma abordadem *parameter driven*; e os modelos ZIP e ZINB autorregressivos em uma abordagem *observation driven*. Os modelos foram aplicados em séries de contagem de casos de sífilis nos Estados Unidos.

Utilizando a abordagem bayesiana, Andrade, Andrade e Ehlers (2016) estudaram modelos GARMA com as distribuições Poisson, binomial e binomial negativa.

Em relação ao uso da distribuição COM-Poisson no contexto de séries temporais, Sunecher, Khan e Jowaheer (2018) desenvolveram um modelo autorregressivo não estacionário de valores inteiros (INAR).

Até o momento, não foram publicados estudos que sejam de nosso conhecimento sobre modelos bayesianos para séries temporais aplicados em dados zero-modificados utilizando o modelo COM-Poisson.

1.5 Estrutura da tese

A tese[1] é dividida da seguinte maneira. No capítulo 2 são apresentados os modelos da família série de potência ARMA zero-modificados (ZM), sendo definidos ainda os modelos Poisson ARMA ZM e COM-Poisson ARMA ZM.

O capítulo 3 ilustra como o método bayesiano, assim como ferramentas de comparação de

[1] O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

modelos, são utilizados nos capítulos subsequentes. Um estudo com dados artificiais é realizado no capítulo 4 e dados reais são analisados no capítulo 5.

A previsão para futuras observações é estudada em um capítulo à parte, no capítulo 6. Por fim, a tese se encerra no capítulo 7, que traz considerações finais e possibilidades de trabalhos futuros.

CAPÍTULO

2

MODELOS DA FAMÍLIA SÉRIE DE POTÊNCIA ZM

A família de distribuições série de potência será apresentada neste capítulo, na seção 2.1. Essa família é a base para a construção dos modelos série de potência zero-modificados (ZM). A transição para a nomenclatura no contexto de séries temporais será dada na seção 2.2. Ao final do capítulo serão dados como exemplos os modelos Poisson e Conway-Maxwell-Poisson (COM-Poisson), todos zero-modificados.

2.1 Modelos da família série de potência

Nos modelos de séries temporais com modificação no número de zeros estudados neste trabalho, foi utilizada uma classe de distribuições conhecida como família série de potência (PS). Essa classe é uma família de distribuições que generaliza diversas distribuições discretas comuns, como a Poisson, Poisson generalizada, binomial, binomial negativa, dentre outras. A função massa de probabilidade de uma variável aleatória cuja distribuição pertence à família PS pode ser escrita como:

$$f_{PS}(y;\mu,\varphi) = \frac{a(y,\varphi)g(\mu,\varphi)^y}{f(\mu,\varphi)}, y \in \mathcal{A}_s. \tag{2.1}$$

O suporte da distribuição é dado por $\mathcal{A}_s = \{s, s+1, s+2, \dots\}$ ou $\mathcal{A}_s = \{s, s+1, s+2, \dots s+n\}$, em que s é um número natural positivo. As funções $a(y,\varphi), g(\mu,\varphi)$ e $f(\mu,\varphi) = \sum_{y\in\mathcal{A}_s} a(y,\varphi)g(\mu,\varphi)^y$ são positivas. Além disso, as funções $g(\mu,\varphi)$ e $f(\mu,\varphi)$ são finitas e possuem derivadas de segunda ordem.

Alguns resultados são gerais para as distribuições na família PS, como por exemplo a média e a variância, que são dadas por

$$\mathbb{E}(Y) = \frac{g(\mu,\varphi)f'(\mu,\varphi)}{f(\mu,\varphi)g'(\mu,\varphi)} \qquad \text{e} \qquad \text{Var}(Y) = \frac{g(\mu,\varphi)}{g'(\mu,\varphi)},$$

respectivamente. As provas desses resultados podem ser encontradas em Gupta (1974).

A Tabela 1 mostra algumas distribuições da família PS de acordo com as funções que a identificam.

Tabela 1 – Distribuições da família série de potência

Nome	$a(y,\varphi)$	$g(\mu,\varphi)$	$f(\mu,\varphi)$	$\mathcal{A}_s$
Poisson	$1/y!$	μ	e^{μ}	$\{0,1,2,\dots\}$
Poisson generalizada	$\frac{(1+\varphi y)^{y-1}}{y!}$	$\mu e^{\frac{-\mu\varphi}{1+\mu\varphi}}$	$e^{\mu/(1+\mu\varphi)}$	$\{0,1,2,\dots\}$
Binomial	$\binom{n}{y}$	$\frac{\mu}{n-\mu}$	$(\frac{n}{n-\mu})^n$	$\{0,1,2,\dots,n\}$
Binomial negativa	$\frac{\Gamma(\varphi+y)}{\Gamma(\varphi)y!}$	$\frac{\mu}{1+\varphi}$	$\left(\frac{\varphi}{\mu+\varphi}\right)^{-\varphi}$	$\{0,1,2,\dots\}$
Conway-Maxwell-Poisson	$(1/y!^{\varphi})$	μ^{φ}	$\sum_{n=0}^{\infty}\left(\frac{\mu^n}{n!}\right)^{\varphi}$	$\{0,1,2,\dots\}$

FONTE: Adaptado de Conceição (2013, p. 7).

2.2 Modelos ARMA ZM

Nesta seção será apresentada a forma geral de um modelo série de potência ARMA zero-modificado (PS ARMA ZM). Considere então os dados em um contexto de séries temporais, $Y_t = \{y_t, t \in T\}$, em que T é um conjunto de índices e $\mathcal{F}_{t-1} = (y_{t-1},\dots,y_1,\mu_{t-1},\dots,\mu_1,p_{t-1},\dots,p_1)$ a informação obtida até o tempo t. Assim, a distribuição condicional série de potência zero-modificada (ZMPS) é dada por

$$f_{ZMPS}(y_t;\mu_t,\varphi,p_t|\mathcal{F}_{t-1}) = (1-p_t)1_{(y_t)} + p_t f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}), \tag{2.2}$$

em que $f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) = \frac{a(y_t,\varphi)g(\mu_t,\varphi)^{y_t}}{f(\mu_t,\varphi)}, y_t \in \mathcal{A}_s$. A variável $1_{(y_t)}$ é igual a 1 se $y_t = 0$ e zero caso contrário. Para o parâmetro p_t, tem-se que

$$0 \leq p_t \leq \frac{1}{1 - f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})}.$$

Para diferentes valores de p_t, diferentes modelos surgem em relação ao número de zeros, de acordo com a Tabela 2. Veja Conceição (2013) para provas e detalhes.

Tabela 2 – Distribuições ZMPS e o parâmetro p_t

Valores de p_t	Tipo de distribuição
$p_t = 0$	Distribuição degenerada com toda a massa no ponto zero
$0 < p_t < 1$	Distribuição série de potência zero inflacionada - ZIPS
$p_t = 1$	Distribuição série de potência usual - PS
$1 < p_t < \frac{1}{1-f_{PS}(0;\mu_t,\varphi)}$	Distribuição série de potência zero deflacionada - ZDPS
$p_t = \frac{1}{1-f_{PS}(0;\mu_t,\varphi)}$	Distribuição série de potência zero truncada - ZTPS

Note que podemos reescrever a equação (2.2) como

$$f_{ZMPS}(y_t;\mu_t,\varphi,p_t|\mathcal{F}_{t-1}) = \{1-p_t[1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})]\}1_{(y_t)} + p_t f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1})(1-1_{(y_t)}).$$

Escrevendo $p_t = \omega_t/[1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})]$, tem-se que $\omega_t = p_t[1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})]$, de forma que $\omega_t \in [0,1]$. Com essa nova parametrização, o modelo ZMPS pode ser escrito como um modelo de Hurdle, isto é,

$$f_{ZMPS}(y_t;\mu_t,\varphi,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTPS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}),$$

em que ZTPS representa a distribuição condicional série de potência zero truncada, isto é,

$$f_{ZTPS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) = \frac{f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1})}{1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})}(1-1_{(y_t)}).$$

Da mesma forma que os valores de p_t qualificam uma distribuição em relação à modificação no zero, o parâmetro ω_t também possui essa propriedade. A Tabela 3 mostra o que ocorre com a presença de zeros na distribuição dos dados conforme varia o valor de ω_t.

Tabela 3 – Distribuições ZMPS e o parâmetro ω_t

Valores de ω_t	Tipo de distribuição
$\omega_t = 0$	Distribuição degenerada com toda a massa no ponto zero
$0 < \omega_t < 1 - f_{PS}(0;\mu_t,\varphi)$	Distribuição série de potência zero inflacionada - ZIPS
$\omega_t = 1 - f_{PS}(0;\mu_t,\varphi)$	Distribuição série de potência usual - PS
$1 - f_{PS}(0;\mu_t,\varphi) < \omega_t < 1$	Distribuição série de potência zero deflacionada - ZDPS
$\omega_t = 1$	Distribuição série de potência zero truncada - ZTPS

A média e variância condicionais das distribuições da família série de potência zero-modificadas são dadas por (CONCEIÇÃO, 2013, p. 25):

$$\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = p_t\,\mu_t = \frac{\omega_t\mu_t}{1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})}$$

$$\mathrm{Var}(Y_t|\mathcal{F}_{t-1}) = p_t[\sigma_t^2+(1-p_t)\mu_t^2] = \frac{\omega_t[\sigma_t^2+(1-p_t)\mu_t^2]}{1-f_{PS}(0;\mu_t,\varphi|\mathcal{F}_{t-1})},$$

em que μ_t e σ_t^2 são, respectivamente, a média e variância da distribuição série de potência associada.

Podemos então escrever o modelo série de potência ARMA zero-modificado (PS ARMA ZM) como sendo:

$$\begin{aligned} &Y_t|\mathcal{F}_{t-1} \sim ZMPS(\mu_t,\varphi,\omega_t) \\ &\log(\mu_t) = \boldsymbol{x}_t'\boldsymbol{\beta} + \sum_{j=1}^{r}\phi_j\log\left(y_{t-j}^{\star}-\boldsymbol{x}_{t-j}'\boldsymbol{\beta}\right) + \sum_{j=1}^{q}\theta_j\log\left(y_{t-j}^{\star}/\mu_{t-j}\right) \\ &\mathrm{logit}(\omega_t) = \boldsymbol{z}_t'\boldsymbol{\gamma} + \sum_{j=1}^{r^*}\delta_j\log\left(y_{t-j}^{\star}-\boldsymbol{z}_{t-j}'\boldsymbol{\gamma}\right) + \sum_{j=1}^{q^*}\nu_j\log\left(y_{t-j}^{\star}/\mu_{t-j}\right), \end{aligned} \tag{2.3}$$

em que $\boldsymbol{x}_t' = (x_{t,1},\ldots,x_{t,u})$ e $\boldsymbol{z}_t' = (z_{t,1},\ldots,z_{t,v})$ são vetores de variáveis explicativas, e os coeficientes regressores para as partes loglinear e logística são, respectivamente, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_u)'$ e $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_v)'$. Nesta tese, para simplificação, os vetores $\boldsymbol{\phi}$ e $\boldsymbol{\delta}$ são considerados de mesmo tamanho e a última soma é omitida na função de ligação logit.

Note que $\log(y_{t-j})$ não possui valor definido quando y_{t-j} é igual a zero. Para contornar esse problema, utilizando a recomendação de Zeger e Qaqish (1988), define-se $y_{t-j}^{\star} = \max(y_{t-j},c)$, em que $0 < c < 1$.

Outras funções de ligação para ω_t poderiam ser utilizadas, como probit ou complemento log-log, descritas em Montgomery, Peck e Vining (2006).

2.2.1 Índice de dispersão

Alguns índices são de interesse para quantificar o quão sub ou sobredispersos estão os dados. Para isso há o índice de dispersão, estudado por Cox e Lewis (1966), que é dado pela seguinte expressão:

$$D_{ZMPS_t} = \frac{\text{Var}(Y_t|\mathcal{F}_{t-1})}{\mathbb{E}(Y_t|\mathcal{F}_{t-1})} = \frac{p_t[\sigma_t^2 + (1-p_t)\mu_t^2]}{p_t\mu_t} = \frac{\sigma_t^2}{\mu_t} + (1-p_t)\mu_t, \tag{2.4}$$

isto é, a razão entre a variância e a média condicionais. Quando $D_{ZMPS_t} < 1$ os dados são subdispersos e quando $D_{ZMPS_t} > 1$ há sobredispersão. No caso de $D_{ZMPS_t} = 1$, tem-se uma distribuição condicional equidispersa, que é o caso da distribuição de Poisson.

Observe na equação (2.4) que o índice de dispersão da distribuição condicional série de potência zero-modificada traz em sua fórmula o próprio índice de dispersão da distribuição série de potência usual, σ_t^2/μ_t.

Como μ_t é positivo, o índice D_{ZMPS_t} indicará sub ou sobredispersão em relação à distribuição PS usual dependendo do valor do parâmetro p_t.

Note que se p_t estiver entre 0 e 1, então $D_{ZMPS_t} = D_{PS_t} + \alpha_t\mu_t$, em que α_t é um valor positivo e D_{PS_t} o índice de dispersão da distribuição PS usual. Assim, nesse caso há uma sobredispersão em relação à distribuição PS original.

No caso de $1 < p_t \leq 1/[1 - f_{PS}(0;\mu_t,\varphi)]$, tem-se que $D_{ZMPS_t} = D_{PS_t} + \alpha_t\mu_t$, em que α_t é um número negativo. Ou seja, há uma subdispersão em relação à distribuição original.

Na seção 2.3 são mostrados os índices de dispersão para as distribuições estudadas.

2.3 Alguns modelos ARMA ZM

Nas próximas subseções, são vistos mais especificadamente os modelos estudados neste trabalho.

2.3.1 Modelo Poisson ARMA ZM

A distribuição condicional de Poisson pode ser escrita na família Série de Potência como

$$f_P(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!} = \frac{(1/y_t!)(\mu_t)^{y_t}}{\sum_{n=0}^{\infty}\mu_t^n/n!}, \quad y_t \in \mathcal{A}_0.$$

Note que o parâmetro φ da família PS é igual a 1 no caso da distribuição de Poisson. Assim, usando a equação (2.1), a distribuição de Poisson pode ser escrita na família PS com $a(y_t) = 1/y_t!$, $g(\mu_t) = \mu_t$ e $f(\mu_t) = e^{\mu_t} = \sum_{n=0}^{\infty}\mu_t^n/n!$.

A média e variância condicionais para o modelo Poisson ARMA, i.e., quando $p_t = 1$, são dadas por $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \text{Var}(Y_t|\mathcal{F}_{t-1}) = \mu_t$.

O modelo Poisson zero-modificado pode ser escrito como

$$f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}), \tag{2.5}$$

em que ZTP representa a distribuição condicional de Poisson zero truncada, isto é,

$$f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{f_P(y_t;\mu_t|\mathcal{F}_{t-1})}{1-f_P(0;\mu_t|\mathcal{F}_{t-1})}(1-1_{(y_t)}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!(1-e^{-\mu_t})}(1-1_{(y_t)}).$$

As funções de ligação para os parâmetros μ_t e ω_t são dadas na equação (2.3). O modelo Poisson zero modificado foi estudado por Conceição, Andrade e Louzada (2014), porém nesse estudo as observações dos conjuntos de dados analisados foram consideradas independentes, sem haver uma estrutura de autocorrelação nos modelos ajustados.

Usando as equações da média e variância para distribuições da família série de potência zero modificadas, encontradas em Conceição (2013, p. 25), temos que para o modelo Poisson ARMA ZM a média e variância condicionais são dadas por

$$\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = p_t\mu_t = \frac{\omega_t\mu_t}{1-f_P(0;\mu_t|\mathcal{F}_{t-1})} = \frac{\omega_t\mu_t}{1-e^{-\mu_t}} \text{ e}$$

$$\begin{aligned}\mathrm{Var}(Y_t|\mathcal{F}_{t-1}) &= p_t[\sigma_t^2 + (1-p_t)\mu_t^2] \\ &= \frac{\omega_t}{1-e^{-\mu_t}}\left[\mu_t + \left(1-\frac{\omega_t}{1-e^{-\mu_t}}\right)\mu_t^2\right] \\ &= \frac{\omega_t\mu_t[(1-e^{-\mu_t})(1+\mu_t)-\omega_t\mu_t]}{(1-e^{-\mu_t})^2},\end{aligned}$$

em que μ_t e σ_t^2 são a média e variância da distribuição de Poisson. Como na distribuição de Poisson a média é igual à variância, temos que $\mu_t = \sigma_t^2$. No modelo Poisson ARMA ZM, a média e a variância podem assumir valores diferentes, o que torna esse modelo mais flexível que o modelo Poisson ARMA usual.

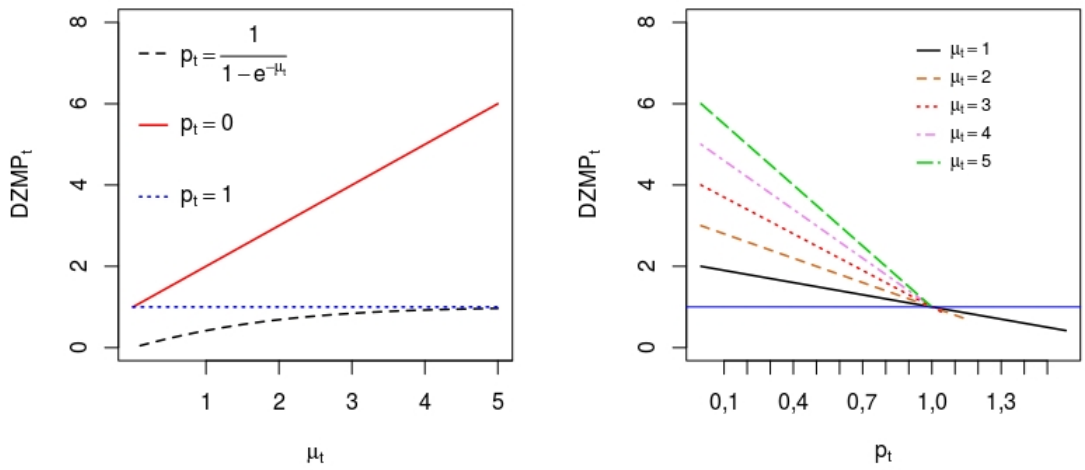
O índice de dispersão para o modelo Poisson ZM é dado por:

$$D_{ZMP_t} = \frac{\mathrm{Var}(Y_t|\mathcal{F}_{t-1})}{\mathbb{E}(Y_t|\mathcal{F}_{t-1})} = \frac{\sigma_t^2}{\mu_t} + (1-p_t)\mu_t = 1 + (1-p_t)\mu_t.$$

Assim, se $0 < p_t < 1$, que implica zero-inflação, tem-se sobredispersão; e se $1 < p_t < \frac{1}{1-e^{-\mu_t}}$, implicando zero-deflação, tem-se em subdispersão. Como notou Conceição (2013, p. 31), essa

característica ocorre com a distribuição de Poisson zero-modificada: zero-inflação implica sobredispersão e zero-deflação implica subdispersão. Para ilustrar o comportamento do índice de dispersão em relação aos parâmetros μ_t e p_t, oberve a Figura 1.

Figura 1 – Índice de dispersão em função de μ_t e p_t para a distribuição de Poisson zero-modificada



A região entre as curvas em que $p_t = 0$ e $p_t = 1$ é a região de inflação de zeros, e é também a região em que o índice de dispersão é maior que 1, isto é, em que há sobredispersão. Da mesma forma, a região entre as curvas em que $p_t = 1$ e $p_t = 1/1 - e^{-\mu_t}$ é a região em que a distribuição é zero-deflacionada, e também é subdispersa. Isso explica visualmente o comentário dado no parágrafo anterior.

É importante notar que como μ_t e p_t variam ao longo do tempo, é possível que os modelos estudados neste trabalho permitam que os dados sejam em um período de tempo zero-inflacionados e em outro período de tempo zero-deflacionados. No entanto, para simplificar a visualização de como o índice de dispersão é influenciado pelos valores de μ_t e p_t, um dos parâmetros foi fixado para cada gráfico.

No primeiro gráfico da Figura 1, pode-se ver como D_{ZMP_t} varia conforme μ_t aumenta, com p_t fixado em três valores: $p_t = 0$, $p_t = 1$ e $p_t = \frac{1}{1-e^{-\mu_t}}$. Quando $p_t = 1$, tem-se que $D_{ZMP_t} = 1 + (1-1)\mu_t = 1$, isto é, há equidispersão. E quando $p_t = \frac{1}{1-e^{-\mu_t}}$, $D_{ZMP_t} = 1 + \left(1 - \frac{1}{1-e^{-\mu_t}}\right)\mu_t$. Assim, $D_{ZMP_t} < 1$ e, portanto, há subdispersão.

No segundo gráfico da Figura 1, D_{ZMP_t} varia no intervalo de possíveis valores de p_t, para alguns valores de μ_t fixados. Por exemplo, quando $\mu_t = 1$, D_{ZMP_t} começa valendo 2 quando $p_t = 0$ – caracterizando sobredispersão – e decresce até chegar ao valor de $0{,}418$, ou seja, ultrapassa a reta indicando equidispersão, quando $D_{ZMP_t} = 1$, no ponto $p_t = 1$ e então chega à região de subdispersão.

O mesmo comportamento ocorre quando $\mu_t = 3$, porém os valores mínimo e máximo que D_{ZMP_t} assume nesse caso são $0,843$ e 4, respectivamente. Quando $\mu_t = 5$, D_{ZMP_t} começa valendo 6 para $p_t = 0$ e termina valendo $0,966$ para $p_t = \frac{1}{1-e^{-5}} = 1,007$.

Sumarizando, quando μ_t é fixado, D_{ZMP_t} relaciona-se com p_t através do segmento de reta com coeficientes linear e angular iguais a $(1+\mu_t)$ e $-\mu_t$, respectivamente.

2.3.2 Modelo COM-Poisson ARMA ZM

A distribuição Conway-Maxwell-Poisson, também conhecida como COM-Poisson, foi proposta pela primeira vez por Conway e Maxwell (1962) no contexto de teoria das filas. Apesar de pouco utilizada inicialmente, a partir de 2005, com o estudo de Shmueli et al. (2005), a distribuição COM-Poisson passou a ser amplamente estudada, por estatísticos e não-estatísticos de diversas áreas.

Essa distribuição passou a ser muito apreciada pelo fato de possuir um parâmetro que controla a dispersão dos dados, o que possibilita tanto a subdispersão quanto a sobredispersão.

Sellers, Borle e Shmueli (2012) sintetizaram os métodos e aplicações relacionados ao estudo da distribuição COM-Poisson. As aplicações se estendem desde a linguística até o comércio, passando por transporte e biologia.

Uma reparametrização da distribuição COM-Poisson, que permite uma melhor interpretação dos parâmetros, foi estudada por Guikema e Goffelt (2008). Com essa nova parametrização, a distribuição condicional COM-Poisson pode ser escrita na família série de potência da seguinte forma:

$$f_{CP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) = \frac{\frac{1}{y_t!^{\varphi}}(\mu_t^{\varphi})^{y_t}}{\sum_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}}, \quad y_t \in \mathcal{A}_0. \tag{2.6}$$

Comparando a distribuição dada pela equação (2.6) com a equação (2.1), a distribuição COM-Poisson possui as seguintes funções que a classificam como uma distribuição pertencente à família série de potência: $a(y_t,\varphi) = 1/y_t!^{\varphi}$, $g(\mu_t,\varphi) = \mu_t^{\varphi}$ e $f(\mu_t,\varphi) = \sum_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}$.

O parâmetro φ é chamado de parâmetro de dispersão, uma vez que quando $\varphi < 1$ há sobredispersão e quando $\varphi > 1$ há subdispersão. Note que quando o parâmetro φ é igual a 1, a distribuição COM-Poisson se reduz à distribuição de Poisson, sendo esta, portanto, um caso particular.

Outros casos particulares são a distribuição geométrica e a distribuição de Bernoulli. É possível identificar a distribuição geométrica ao se fazer $\varphi = 0$ na parametrização em que $\lambda_t = \mu_t^{\varphi}$,

para valores de λ_t menores que 1. Já a distribuição de Bernoulli é obtida no limite quando $\varphi \to \infty$, e seu parâmetro é $\lambda_t/(1+\lambda_t)$.

A média e variância condicionais da distribuição COM-Poisson são dados por $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = \frac{1}{\varphi}\frac{\partial \log f(\mu_t,\varphi)}{\partial \log \mu_t}$ e $\text{Var}(Y_t|\mathcal{F}_{t-1}) = \frac{1}{\varphi^2}\frac{\partial^2 \log f(\mu_t,\varphi)}{\partial \log^2 \mu_t}$, respectivamente. Note que esses sumários da distribuição COM-Poisson não possuem uma expressão fechada devido à função normalizadora $f(\mu_t,\varphi)$. No entanto, expressões assintóticas podem ser obtidas em função dos parâmetros (GUIKEMA; GOFFELT, 2008):

$$\mu_t^* = \mathbb{E}(Y_t|\mathcal{F}_{t-1}) \approx \mu_t + \frac{1}{2\varphi} - \frac{1}{2} \tag{2.7}$$

$$\sigma_t^2 = \text{Var}(Y_t|\mathcal{F}_{t-1}) \approx \frac{\mu_t}{\varphi} \tag{2.8}$$

Segundo Guikema e Goffelt (2008), essas aproximações são precisas quando $\mu_t > 10$. O modelo COM-Poisson zero-modificado pode ser escrito como:

$$f_{ZMCP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTCP}(y_t;\mu_t|\mathcal{F}_{t-1}),$$

em que ZTCP representa a distribuição condicional COM-Poisson zero truncada, isto é,

$$\begin{aligned} f_{ZTCP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) &= \frac{f_{CP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1})}{1-f_{CP}(0;\mu_t,\varphi|\mathcal{F}_{t-1})}(1-1_{(y_t)}) \\ &= \frac{1/y_t!^{\varphi}\mu_t^{\varphi y_t}}{\sum_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}-1}(1-1_{(y_t)}). \end{aligned}$$

As funções de ligação para os parâmetros μ_t e ω_t são dadas na equação (2.3). Utilizando as equações (2.7) e (2.8), a média e a variância condicionais do modelo COM-Poisson ARMA ZM são aproximadas pelas expressões:

$$\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = p_t\mu_t^* = \frac{\omega_t\mu_t^*}{1-f_{CP}(0;\mu_t,\varphi)} \approx \frac{\omega_t(\mu_t+1/2\varphi-1/2)}{1-1/f(\mu_t,\varphi)} \text{ e}$$

$$\begin{aligned}\text{Var}(Y_t|\mathcal{F}_{t-1}) &= p_t[\sigma_t^2+(1-p_t)\mu_t^{*2}]\\ &\approx \frac{\omega_t}{1-1/f(\mu_t,\varphi)}\left[\frac{\mu_t}{\varphi}+\left(1-\frac{\omega_t}{1-1/f(\mu_t,\varphi)}\right)\left(\mu_t+\frac{1}{2\varphi}-\frac{1}{2}\right)^2\right]\\ &\approx \frac{\omega_t}{\varphi[(f(\mu_t,\varphi)-1)/f(\mu_t,\varphi)]^2}\times\\ &\quad\left[\mu_t\left(\frac{f(\mu_t,\varphi)-1}{f(\mu_t,\varphi)}\right)+\varphi\left(\mu_t+\frac{1}{2\varphi}-\frac{1}{2}\right)^2\left(\frac{f(\mu_t,\varphi)-1}{f(\mu_t,\varphi)}-\omega_t\right)\right],\end{aligned}$$

em que μ_t^* e σ_t^2 são a média e a variância condicionais do modelo COM-Poisson e $f(\mu_t,\varphi) = \sum_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}$, a constante normalizadora. O índice de dispersão para o modelo COM-Poisson ZM é dado por:

$$\begin{aligned}D_{ZMCP} &= \frac{\text{Var}(Y_t|\mathcal{F}_{t-1})}{\mathbb{E}(Y_t|\mathcal{F}_{t-1})} = \frac{\sigma_t^2}{\mu_t^*}+(1-p_t)\mu_t^*\\ &\approx \frac{\mu_t/\varphi+(1-p_t)(\mu_t+1/2\varphi-1/2)^2}{\mu_t+1/2\varphi-1/2}.\end{aligned}$$

A Figura 2 mostra o índice de dispersão para a distribuição COM-Poisson, com $\varphi = 0,5$. Note que, usando $\varphi = 0,5$, a região de inflação de zeros é também uma região em que o índice de dispersão é maior que 1, ou seja, em que há sobredispersão. No entanto, não é possível afirmar que a região de deflação de zeros implique subdispersão. Apesar de haver uma pequena sobredispersão para valores crescentes de μ_t, ela ocorre.

Figura 2 – Índice de dispersão em função de μ_t e p_t para a distribuição COM-Poisson zero-modificada, com $\varphi = 0,5$

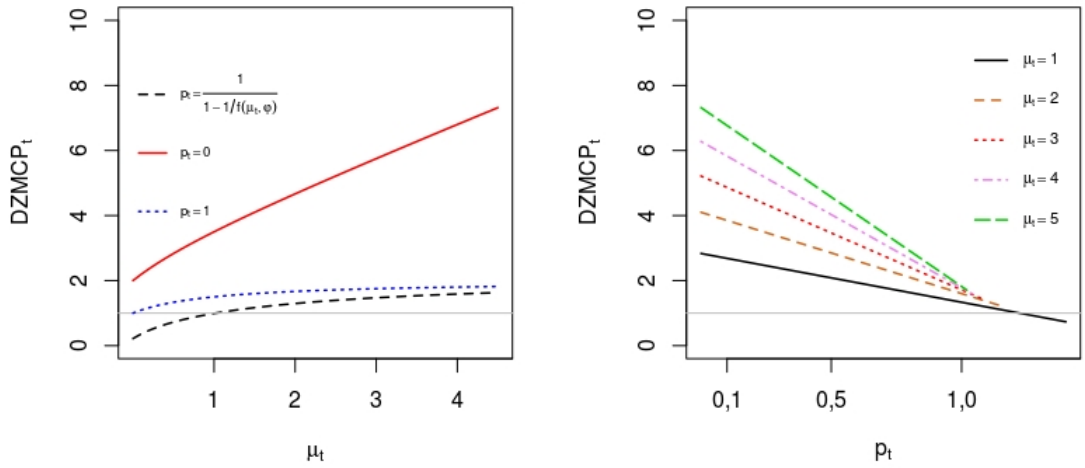
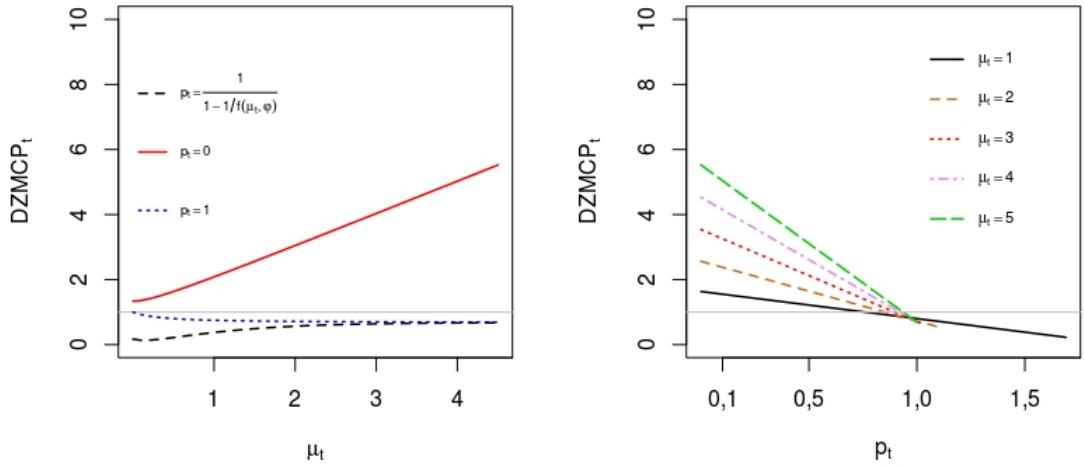


Figura 3 – Índice de dispersão em função de μ_t e p_t para a distribuição COM-Poisson zero-modificada, com $\varphi = 1,5$



As conclusões se invertem quando fazemos os gráficos com $\varphi = 1,5$, permitindo uma maior subdispersão dos dados. Observe que, na Figura 3, a região de deflação de zeros também é de subdispersão, embora a região de inflação de zeros não seja sempre de sobredispersão.

2.4 Simulação de valores aleatórios

Para gerar valores aleatórios de distribuições série de potência ARMA zero-modificadas, o método da inversa é utilizado. Note que, como há termos autorregressivos e de médias móveis

nas funções de ligação, para gerar uma amostra do modelo será necessário prover valores iniciais para que o algoritmo consiga gerar os primeiros valores aleatórios.

Dessa maneira, munimos o algoritmo com m valores iniciais para a amostra, que denotamos por $\boldsymbol{y_m} = (y_1, \ldots, y_m)$. Assim, o programa é capaz de calcular internamente $\log(y_t^\star)$, para $t = 1, \ldots, i$, em que $i = \max(r, q)$ e $m \geq i$.

O método da transformada inversa pode ser usado para gerar uma amostra aleatória da distribuição pertencente à família série de potência do modelo PS ARMA ZM. Tal método se baseia no seguinte: seja uma função de distribuição acumulada (f.d.a.) $F(y)$, em que $y \in R$. Considere ainda a função inversa $F^{-1}(u)$, em que $u \in [0, 1]$. Defina $Y = F^{-1}(U)$, em que $U \sim Unif(0, 1)$. Então Y tem distribuição $F(y)$, i.e., $P(Y \leq y) = F(y)$.

Esse método permite gerar valores de uma distribuição de interesse, também chamada de distribuição alvo, com um gerador de números aleatórios da distribuição $Unif(0, 1)$ e a f.d.a. da variável aleatória (v.a.) de interesse.

Para uma v.a. discreta, como é o caso de interesse neste trabalho, o método da transformada inversa é executado com o seguinte algoritmo:

Algoritmo 1 – Método da transformada inversa

Simule u de uma v.a. $Unif(0, 1)$
se $u < P(Y = 0)$, **então**
    defina $Y = 0$
senão
    se $\sum_{i=0}^{k-1} P(Y = i) < u \leq \sum_{i=0}^{k} P(Y = i)$, **então**
        defina $Y = k$
    fim se
fim se
retorna Y

No caso especial das distribuições da família série de potência, há a fórmula recursiva (CONSUL; FAMOYE, 2006) que pode ser usada para otimizar o algoritmo, dada por:

$$f_{ZMPS}(y_t; \mu_t, \varphi, p_t | \mathcal{F}_{t-1}) = \frac{a(y_t, \varphi) g(\mu_t, \varphi)}{a(y_t - 1, \varphi)} f_{ZMPS}(y_t - 1; \mu_t, \varphi, p_t | \mathcal{F}_{t-1}), \text{ para } y_t \geq 2.$$

Para chegar ao resultado, note que ao dividirmos a função massa de probabilidade (f.m.p.) no ponto y_t pela f.m.p. no ponto $y_t - 1$, obtemos

$$\frac{f_{ZMPS}(y_t; \mu_t, \varphi, p_t | \mathcal{F}_{t-1})}{f_{ZMPS}(y_t - 1; \mu_t, \varphi, p_t | \mathcal{F}_{t-1})} = \frac{(1 - p_t) 1_{(y_t)} + p_t f_{PS}(y_t; \mu_t, \varphi | \mathcal{F}_{t-1})}{(1 - p_t) 1_{(y_t - 1)} + p_t f_{PS}(y_t - 1; \mu_t, \varphi | \mathcal{F}_{t-1})}, \text{ para } y_t \geq 2. \quad (2.9)$$

Como a equação (2.9) vale para $y_t \geq 2$, tem-se que as funções indicadoras serão iguais a zero. Logo, a equação (2.9) se reduz a

$$\frac{f_{ZMPS}(y_t;\mu_t,\varphi,p_t|\mathcal{F}_{t-1})}{f_{ZMPS}(y_t-1;\mu_t,\varphi,p_t|\mathcal{F}_{t-1})} = \frac{f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1})}{f_{PS}(y_t-1;\mu_t,\varphi|\mathcal{F}_{t-1})}, \text{ para } y_t \geq 2. \tag{2.10}$$

Mas

$$\begin{aligned}\frac{f_{PS}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1})}{f_{PS}(y_t-1;\mu_t,\varphi|\mathcal{F}_{t-1})} &= \frac{a(y_t,\varphi)g(\mu_t,\varphi)^{y_t}}{f(\mu_t,\varphi)} \times \frac{f(\mu_t,\varphi)}{a(y_t-1,\varphi)g(\mu_t,\varphi)^{y_t-1}} \\ &= \frac{a(y_t,\varphi)g(\mu_t,\varphi)}{a(y_t-1,\varphi)}, \text{ para } y_t \geq 2.\end{aligned}$$

Pela equação (2.10), tem-se que

$$\begin{aligned}&\frac{f_{ZMPS}(y_t;\mu_t,\varphi,p_t|\mathcal{F}_{t-1})}{f_{ZMPS}(y_t-1;\mu_t,\varphi,p_t|\mathcal{F}_{t-1})} = \frac{a(y_t,\varphi)g(\mu_t,\varphi)}{a(y_t-1,\varphi)} \\ &\Rightarrow f_{ZMPS}(y_t;\mu_t,\varphi,p_t|\mathcal{F}_{t-1}) = \frac{a(y_t,\varphi)g(\mu_t,\varphi)}{a(y_t-1,\varphi)} f_{ZMPS}(y_t-1;\mu_t,\varphi,p_t|\mathcal{F}_{t-1}), \text{ para } y_t \geq 2.\end{aligned}$$

Conceição (2013, p. 37) descreve o algoritmo para geração de valores aleatórios de v.a.'s com distribuições da família série de potência zero-modificadas. Neste trabalho, utilizou-se como base o algoritmo descrito em Conceição (2013, p. 37), com as mudanças necessárias para incluir as funções de ligação de μ_t e ω_t e, como comentado anteriormente neste capítulo, podendo utilizar valores iniciais para dar início ao algoritmo.

Para que os valores inciais não exerçam influência, os primeiros valores podem ser descartados e aproveitamos apenas a amostra resultante desconsiderando 100 valores inicias.

O pseudo-código do algoritmo gerador de números aleatórios do modelo PS ARMA ZM está descrito a seguir, considerando que foi informado em tal função o seguinte:

- O valor de n, quantidade de valores aleatórios a serem gerados;
- Os vetores com valores dos parâmetros $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\gamma}$ e $\boldsymbol{\delta}$, com tamanhos u, r, q, v e r, respectivamente;
- Os vetores de variáveis explicativas $\boldsymbol{x}_t$ e $\boldsymbol{z}_t$, com tamanhos u e v, respectivamente;
- O vetor $\boldsymbol{y}_t$, de observações passadas, com tamanho $\max(r,q)$;
- O vetor de médias iniciais $\boldsymbol{\mu_0}$, com tamanho q.

Algoritmo 2 – Algoritmo gerador de observações do modelo PS ARMA ZM

procedimento ZMPS
  Crie o vetor $zmps$ para armazenar os valores aleatórios
  Declare as funções $a(y_t, \varphi)$, $g(\mu_t, \varphi)$ e $f(\mu_t, \varphi)$ conforme a distribuição PS de interesse
  Defina r e q como os tamanhos dos vetores $\boldsymbol{\phi}$ e $\boldsymbol{\theta}$, respectivamente
  Declare N como $n+100$, em que N é o valor de observações geradas
  para t de $m+1$ a N **faça**
    Substitua os valores iguais a zero do vetor $\boldsymbol{y}_t$ por $0,5$, chamando esse novo vetor de $\boldsymbol{y}_t^*$.
    $\mu_t = \exp\left\{\boldsymbol{x}_t'\boldsymbol{\beta} + \sum_{j=1}^{r} \phi_j \log\left(y_{t-j}^{\star} - \boldsymbol{x}_{t-j}'\boldsymbol{\beta}\right) + \sum_{j=1}^{q} \theta_j \log\left(y_{t-j}^{\star}/\mu_{t-j}\right)\right\}$
    $k_t = \exp\left\{\boldsymbol{z}_t'\boldsymbol{\gamma} + \sum_{j=1}^{r} \delta_j \log\left(y_{t-j}^{\star} - \boldsymbol{z}_{t-j}'\boldsymbol{\gamma}\right)\right\}$
    $\omega_t = k_t/(1+k_t)$
    $p_t = \omega_t/[1 - f_{PS}(0; \mu_t, \varphi|\mathcal{F}_{t-1})]$
    $y_t = 0$
    $p_0 = (1-p_t) + \frac{p_t}{f(\mu_t,\varphi)}$
    Gere u de uma distribuição $Unif(0,1)$: $u = runif(1)$
    Declare $F_{y_t} = p_0$
    enquanto $u > F_{y_t}$ **faça**
      $y_t = y_t + 1$
      Calcular $a(y_t, \varphi)$
      se $y_t = 1$ **então**
        $f_{ZMPS}(y_t; \mu_t, \varphi, p_t|\mathcal{F}_{t-1}) = p_t \frac{a(y_t,\varphi)g(\mu_t,\varphi)^{y_t}}{f(\mu_t,\varphi)}$
      senão $f_{ZMPS}(y_t; \mu_t, \varphi, p_t|\mathcal{F}_{t-1}) = \frac{a(y_t,\varphi)g(\mu_t,\varphi)^{y_t}}{a(y_t-1,\varphi)} f_{ZMPS}(y_t - 1; \mu_t, \varphi, p_t|\mathcal{F}_{t-1})$
      fim se
      $F_{y_t+1} = F_{y_t} + f_{ZMPS}(y_t; \mu_t, \varphi, p_t|\mathcal{F}_{t-1})$
    fim enquanto
    Acrescente o resultado no vetor de números aleatórios: $zmps[i] = y_t$
    Atualize os outros vetores: $\boldsymbol{y}_t = c(\boldsymbol{y}_t', y_t)$ e $\boldsymbol{\mu}_0 = c(\boldsymbol{\mu}_0', \mu_t)$
  fim para
  retorna o vetor $zmps$ sem os primeiros 100 valores
fim procedimento

No próximo capítulo, é apresentada a função de verossimilhança parcial para os modelos PS ARMA ZM, assim como a abordagem bayesiana na implementação dos modelos de interesse.

CAPÍTULO

3

MÉTODOS BAYESIANOS EM MODELOS SÉRIE DE POTÊNCIA ARMA ZM

Neste capítulo é apresentada a função de verossimilhança parcial dos modelos série de potência ARMA zero-modificados, assim como o ajuste dos modelos por meio de métodos de Monte Carlo via Cadeias de Markov (MCMC).

3.1 Função de verossimilhança parcial

A função de verossimilhança parcial, como descrita em Kedem e Fokianos (2002, p. 4), será utilizada para implementação do modelo na abordagem bayesiana, juntamente com as distribuições a priori. Segundo Kedem e Fokianos (2002, p. 4), apenas o que já é conhecido até o tempo de observação t é considerado pela função de verossimilhança parcial.

Considere $\boldsymbol{\Theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\theta}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$ o vetor de parâmetros dos modelos PS ARMA ZM. Usando notação semelhante à de Benjamin, Rigby e Stasinopoulos (2003), a função de verossimilhança dos dados $(y_{m+1}, \ldots, y_n)$ condicionados às primeiras observações $(y_1, \ldots, y_m)$, às variáveis $\boldsymbol{x}_t$ e $\boldsymbol{z}_t$ e em $\log(y_t^\star)$, para $t = 1, \ldots, i$, em que $i = \max(r, q)$ e $m \geq i$ é dada por:

$$\begin{aligned} \mathrm{PL}(\boldsymbol{\Theta}; \boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t) &= \prod_{t=m+1}^{n} f_{ZMPS}(y_t, \boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\Theta} | \mathcal{F}_{t-1}) \\ &= \prod_{t=m+1}^{n} (1-\omega_t) 1_{(y_t)} + \omega_t f_{ZTPS}(y_t; \mu_t, \varphi | \mathcal{F}_{t-1}) \end{aligned} \tag{3.1}$$

A função de verossimilhança parcial, apresentada na equação (3.1), não é tratável analiticamente. Por meio da abordagem clássica, poderiam ser utilizados métodos iterativos, como o BFGS, para obter as estimativas de máxima verossimilhança. Neste trabalho, entretanto, optou-se por utilizar a abordagem bayesiana por uma série de motivos, entre eles está a facilidade

de implementação do modelo quando não há solução analítica da maximização da função de verossimilhança.

Entre os motivos mais utilizados para implementação do pensamento bayesiano estão:

a) a possibilidade de incorporação de informação a priori no modelo, seja por meio de estudos anteriores, seja por meio de opiniões de especialistas;

b) quando não há informações anteriores, pode-se utilizar prioris vagas, com variâncias grandes, o que faz com que os dados tenham maior peso em relação à priori na distribuição a posteriori;

c) com apenas os dados do problema e a função de verossimilhança parcial, é possível implementar métodos MCMC;

d) a interpretação das estimativas é direta na abordagem bayesiana. Por exemplo, um intervalo de credibilidade de 95% para um parâmetro nos dá o intervalo no qual a probabilidade de este parâmetro estar contido no intervalo é de 95%.

Assim, a abordagem bayesiana será utilizada neste trabalho. Considere novamente o vetor de parâmetros $\boldsymbol{\Theta} = (\boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\theta}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$. Na abordagem bayesiana, a distribuição a posteriori é obtida como uma combinação entre a distribuição a priori e a verossimilhança, que contém a informação dos dados. Pode-se pensar que a distribuição a posteriori é resultado dessas duas fontes de informação: priori e verossimilhança. Por meio do teorema de Bayes, essa relação pode ser escrita da seguinte forma:

$$\begin{aligned} p(\boldsymbol{\Theta}|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t) &= \frac{\mathrm{PL}(\boldsymbol{\Theta}|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t) \times p(\boldsymbol{\Theta})}{P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t)}, \\ &\propto \mathrm{PL}(\boldsymbol{\Theta}|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t) \times p(\boldsymbol{\Theta}), \end{aligned}$$

em que $p(\boldsymbol{\Theta}|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t)$ é a distribuição a posteriori, $\mathrm{PL}(\boldsymbol{\Theta}|\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t)$ é a função de verossimilhança parcial, $\boldsymbol{x}_t, \boldsymbol{z}_t$ são variáveis exógenas e $p(\boldsymbol{\Theta})$ é a distribuição a priori.

3.2 Distribuições a priori

Até relativamente pouco tempo atrás, a abordagem bayesiana era computacionalmente inviável devido à impossibilidade de se calcular a constante normalizadora $P(\boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t)$ para modelos mais flexíveis e complexos, e apenas alguns modelos, como os conjugados, eram possíveis de serem analisados sem necessidade de recorrer a métodos computacionalmente intensivos.

Com o advento dos métodos MCMC, foi possível amostrar valores da distribuição a posteriori sem a necessidade de se calcular a constante normalizadora e isso permitiu que a

inferência bayesiana fosse aplicada a problemas mais complexos, como por exemplo as aplicações estudadas nesta tese.

Assim, apenas com a verossimilhança parcial, apresentada na seção 3.1, e as distribuições a priori selecionadas para os parâmetros, é possível obter as distribuições a posteriori dos parâmetros do modelo estudado.

Uma vez que não foram utilizadas informações prévias de outros estudos ou opiniões de especialistas sobre os dados analisados nesta tese, as distribuições a priori aqui escolhidas são amplas, vagas na escala, o que faz com que possuam pouco efeito sobre as distribuições a posteriori, prevalecendo assim a informação providenciada pelos dados.

Dessa maneira, as distribuições a priori para os parâmetros terão variabilidade ampla. A distribuição normal foi escolhida pela facilidade de interpretação de seus parâmetros e pelo fato de os parâmetros dos modelos estudados terem suporte nos números reais. De maneira geral, as distribuições a priori são dadas por:

$$\begin{aligned}
\boldsymbol{\beta} &\sim N_u(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2_{\boldsymbol{\beta}} \boldsymbol{I}_u) \\
\boldsymbol{\gamma} &\sim N_v(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}} \boldsymbol{I}_v) \\
\boldsymbol{\phi} &\sim N_r(\boldsymbol{\mu}_{\boldsymbol{\phi}}, \sigma^2_{\boldsymbol{\phi}} \boldsymbol{I}_r) \\
\boldsymbol{\theta} &\sim N_q(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2_{\boldsymbol{\theta}} \boldsymbol{I}_q) \\
\boldsymbol{\delta} &\sim N_r(\boldsymbol{\mu}_{\boldsymbol{\delta}}, \sigma^2_{\boldsymbol{\delta}} \boldsymbol{I}_r),
\end{aligned}$$

em que N_k refere-se à distribuição normal k-variada e $\boldsymbol{I}_k$ representa a matriz identidade de dimensão $k \times k$. Os hiperparâmetros de média e variância podem ser diferentes para cada parâmetro; no entanto, nesta tese os valores utilizados como média e variância são dados por 0 e 10^5, respectivamente, para todos os parâmetros.

Para as distribuições base que possuem parâmetro de dispersão, como é o caso da distribuição COM-Poisson, é necessário aplicar uma distribuição a priori para o parâmetro φ, em que também optou-se por uma distribuição normal com média 0 e variância 10^5.

3.3 Métodos MCMC

3.3.1 JAGS

O programa JAGS (do inglês *Just another Gibbs samppler*), foi utilizado nesta tese na implementação dos modelos de estudo. Trata-se de um programa desenvolvido por Plummer (2003) e que pode ser utilizado juntamente com o programa R por meio de alguns pacotes, como "runjags" e "rjags".

O método utilizado pelo JAGS é o amostrador de Gibbs, ou *Gibbs sampling*, que foi introduzido por Geman e Geman (1984) e popularizado na comunidade estatística por Gelfand e Smith (1990). Trata-se de um caso especial do algoritmo Metropolis-Hastings, que por sua vez é uma generalização do algoritmo Metropolis. Todos esses algoritmos citados estão dentro do arcabouço dos métodos MCMC. O algoritmo Metropolis foi introduzido em 1953, por Metropolis et al. (1953). O segundo "M" da sigla MCMC significa Markov, e a propriedade markoviana presente nesses métodos significa que o próximo elemento da cadeia gerado depende apenas do estado presente, isto é, do valor atual.

O amostrador de Gibbs é um método MCMC, que permite gerar valores das distribuições a posteriori dos parâmetros de interesse de um modelo. A partir da distribuição a posteriori, diversos estudos podem ser feitos para estimar os parâmetros, como por exemplo criar intervalos de credibilidade para os parâmetros ou estudar correlações entre eles.

3.3.2 Diagnóstico dos métodos MCMC

Uma parte importante da implementação de métodos MCMC é seu diagnóstico. Nesse sentido, é importante verificar se as cadeias geradas convergiram, e se convergiram independentemente do ponto inicial. Para verificar se as cadeias convergiram, dois métodos usuais serão utilizados: inspeção visual da cadeia e cálculo da estatística de Gelman-Rubin (Gelman e Rubin (1992), Brooks e Gelman (1998)), também chamada de *shrink factor*.

A **estatística de Gelman-Rubin** calcula a relação entre a variação combinada (variação que combina a variação dentro de cada cadeia e variação entre cadeias) e a variação dentro das cadeias. O resultado esperado é que essa relação seja igual a 1, pois se as cadeias são amostradas de uma mesma população, a variação combinada deve ser semelhante à variação dentro das cadeias. Dessa forma, pode-se avaliar se a convergência das cadeias geradas foi obtida. Brooks e Gelman (1998) fizeram um melhoramento no cálculo dessa estatística para levar em consideração a variabilidade amostral. Nesta tese é adotado o seguinte critério: caso a estatística de Gelman-Rubin seja inferior a $1,1$, conclui-se concluído que a convergência da cadeia foi atingida.

O **tamanho amostral eficaz**, do inglês *effective sample size* (ESS) é uma medida usada para saber se a amostra gerada pelo método MCMC é de fato representativa da população (no caso, a população é a distribuição a posteriori de um determinado parâmetro de interesse). Como são geradas sucessivas observações do parâmetro de interesse e essas observações são correlacionadas com as anteriores, é necessário checar se ao final da utilização de um método MCMC existe de fato uma amostra representativa da distribuição a posteriori.

O tamanho amostral eficaz é calculado com base nas informações de todas as cadeias geradas. De acordo com Kruschke (2015, p. 184), um tamanho amostral eficaz de no mínimo 10 mil é recomendado para uma estimação apropriada da distribuição a posteriori dos parâmetros.

No entanto, esse valor é baseado na experiência em aplicações práticas, e portanto não deve ser tratado como um requisito necessário. Além disso, se o interesse maior for em estimar o valor mediano dos parâmetros, um tamanho amostral eficaz menor pode ser suficiente, o que pode não ser verdade se a intenção for de estimar com precisão os limites inferior e superior dos intervalos de credibilidade.

3.4 Regiões HDI e ROPE

Há várias formas de se definir intervalos de credibilidade bayesianos. Nesta tese, serão utilizados os **intervalos de alta densidade**, ou HDIs, do inglês *highest density intervals*. Esses intervalos consideram os pontos com maior probabilidade, que estão dentro da região, em contraposição aos pontos fora da região, que possuem menor probabilidade. Os HDIs são definidos da seguinte maneira (BOX; TIAO, 1973, p. 123), (BERNARDO; SMITH, 2000, p. 260):

Definição 1. (Intervalos de alta densidade). Uma região $R \subseteq \Theta$ é chamada de região de alta densidade com probabilidade de (1-α)100% para o parâmetro θ com respeito a $p(\theta)$ se

(i) $P(\theta \in R) = 1 - \alpha$;

(ii) $p(\theta_1) \geq p(\theta_2)$ para todo $\theta_1 \in R$ e $\theta_2 \notin R$, exceto em um subconjunto de Θ com probabilidade zero.

Se $p(\theta)$ for uma distribuição a posteriori (priori, preditiva), chamamos de região de alta densidade a posteriori (a priori, preditiva).

Uma **região de equivalência prática**, do inglês *region of practical equivalence*, ou ROPE, é considerada neste trabalho como uma porção da distribuição de interesse (geralmente a distribuição a posteriori de algum parâmetro), que revela a probabilidade de um parâmetro pertencer a um intervalo arbitrário. A ROPE será utilizada ao se testar se um parâmetro em análise é estatisticamente significante. Por exemplo, se for de interesse testar se um parâmetro pode ser considerado igual a zero, quando o HDI a posteriori rejeitar uma ROPE de $-0,1$ a $0,1$, pode-se dizer que o parâmetro é estatisticamente significante, isto é, diferente de zero. É importante notar que a região é definida de acordo com a magnitude das observações e varia de acordo com o cenário de estudo.

O estudo de regiões de equivalência foi motivado por aplicações clínicas, em que se desejava avaliar se a eficácia de um novo tratamento era equivalente a outro padrão. Em vez de determinar um único ponto na hipótese nula (de que os tratamentos são iguais), é considerada uma região de equivalência, ou zona de indiferença, na qual não se tem comprovação de que o novo tratamento seja melhor ou pior que o tratamento padrão. Assim, em vez de simplesmente

concluir se o novo tratamento é melhor (ou pior), pode-se avaliar o quão melhor (ou pior) ele é em relação ao tratamento padrão. Estudos e livros versando sobre ROPE incluem Kruschke (2011), Kruschke (2015), Carlin e Louis (2009), Freedman, Lowe e Macaskill (1984), Hobbs e Carlin (2007), Spiegelhalter, Freedman e Parmar (1994).

3.5 Comparação de modelos

3.5.1 Critério de informação da deviância: DIC

O critério de informação da deviância, em inglês *deviance information criterion* (DIC), é utilizado para comparar modelos candidatos em uma análise de dados. Essa medida busca encontrar o melhor modelo entre aqueles sendo considerados em um estudo, levando em conta o quão bem o modelo analisado se ajusta aos dados, porém compensando com uma penalização de acordo com sua complexidade. A deviância é definida como

$$D = -2\log \mathrm{PL}(\boldsymbol{\Theta}; \boldsymbol{y}_t, \boldsymbol{x}_t, \boldsymbol{z}_t) = -2\sum_{t=m+1}^{n} \log f_{ZMPS}(y_t, \boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\Theta}|\mathcal{F}_{t-1})$$

O DIC foi proposto por Spiegelhalter et al. (2002), e é definido como

$$\begin{aligned} \mathrm{DIC} &= D(\bar{\boldsymbol{\Theta}}) + 2p_D \\ &= \bar{D} + p_D, \end{aligned}$$

em que $\bar{D} = \mathbb{E}(D)$ é a deviância média a posteriori e $p_D = \bar{D} - D(\bar{\boldsymbol{\Theta}})$ é uma medida para o número efetivo de parâmetros, definida como a diferença entre a deviância média a posteriori e a deviância nas estimativas a posteriori dos parâmetros de interesse (SPIEGELHALTER et al., 2002, p. 584). Essa quantidade é facilmente obtida nos métodos MCMC. Segundo o critério de informação da deviância, o melhor modelo será aquele com menor valor de DIC.

3.5.2 Critério de informação bayesiano esperado: EBIC

O critério de informação bayesiano esperado, em inglês *expected Bayesian information criterion*, é outra medida similar ao DIC. Sua definição, assim como relação com o DIC, são dados por (CARLIN; LOUIS, 2009):

$$\begin{aligned} \mathrm{EBIC} &= \bar{D} + p_D \log n \\ &= \mathrm{DIC} - p_D + p_D \log n \\ &= \mathrm{DIC} + p_D(\log(n) - 1), \end{aligned}$$

com n sendo o número de observações. Também segundo o critério de informação bayesiano esperado, o melhor modelo será aquele com menor valor de EBIC.

3.5.3 Análise preditiva a posteriori

Uma maneira de avaliação do modelo, considerando o ajuste aos dados observados, é realizada por meio da análise preditiva a posteriori. Para Rubin (1984),

> *The applied statistician should avoid models that are contradicted by the data in relevant ways — frequency calculations for hypothetical replications can monitor a model's adequacy and help to suggest more appropriate models.*

Rubin (1984) ressalta, portanto, a importância da verificação do ajuste do modelo aos dados de interesse, evitando-se assim o uso de modelos que contradizem os dados. E uma maneira sugerida pelo autor é comparar frequências dos valores observados e de réplicas do modelo candidato.

O método da análise preditiva a posteriori é como segue. Considere o conjunto de dados observados $\boldsymbol{y}' = (y_1, \ldots, y_n)$. M conjuntos $\boldsymbol{y}_i'^{\text{rep}} = (y_{1,i}^{\text{rep}}, \ldots, y_{n,i}^{\text{rep}}), i = 1 \ldots, M$, chamados de réplicas, são simulados a partir dos parâmetros estimados e as distribuições dos dois vetores são comparadas.

O termo "preditiva" é usado devido à distribuição dos dados não observados. O termo "a posteriori" é usado por ser um método condicionado nos dados observados.

A ideia desse método é que, gerando-se observações hipotéticas do modelo em análise, espera-se obter frequências próximas às frequências dos valores dos dados observados. Por exemplo, se em nosso estudo há 50 zeros (dados observados), esperamos encontrar um número de zeros similar após ajustar um modelo apropriado e gerar valores desse modelo (réplicas).

Nesta tese, quando a análise preditiva a posteriori é realizada, são geradas $M = 20$ réplicas do estudo em questão. Por exemplo, quando um modelo Poisson AR(1) ZM é ajustado a um conjunto de n dados observados, retiraremos 20 amostras dos parâmetros do modelo utilizando para isso suas distribuições a posteriori.

Uma vez com as amostras dos parâmetros estimados, geraremos $\boldsymbol{y}_i'^{\text{rep}} = (y_{1,i}^{\text{rep}}, \ldots, y_{n,i}^{\text{rep}}), i = 1 \ldots, M$, usando a função geradora de valores aleatórios do modelo de interesse, conforme seção 2.4 na página 37. Assim, haverá $M = 20$ réplicas do estudo de interesse.

Com os dados observados $\boldsymbol{y}'$ e as réplicas $\boldsymbol{y}_i'^{\text{rep}} = (y_{1,i}^{\text{rep}}, \ldots, y_{n,i}^{\text{rep}}), i = 1, \ldots, M$, gráficos de barras com as frequências de $\boldsymbol{y}'$ e $\boldsymbol{y}_i'^{\text{rep}}$ são comparados, para $i = 1, \ldots, M$. Se o modelo estiver adequado aos dados, espera-se que as frequências sejam similares.

3.5.4 Previsão

Com o modelo escolhido após as etapas de ajuste aos dados, comparação de modelos e análise preditiva a posteriori, o comportamento ao realizar previsões de futuras observações é verificado.

Para avaliar a previsão, o conjunto de dados é dividido em duas partes. A primeira é chamada de dados de treinamento e a segunda de dados de teste. O modelo é implementado apenas com os dados de treinamento, ficando os dados de teste "de fora". Com o modelo implementado, futuras observações são geradas com base nas estimativas dos parâmetros do modelo. O número de observações a serem previstas é chamado de horizonte.

Nesta tese, optou-se por realizar uma análise de previsão considerando um horizonte de seis observações, o equivalente a um semestre para dados mensais. Como serão analisados dados de saúde, as previsões são de interesse para o uso em estabelecimentos e organizações de saúde, possibilitando políticas públicas para priorizar campanhas de vacinação ou outras formas de prevenção e controle de doenças.

Com as futuras observações geradas, estas são comparadas com os dados de teste. Essa comparação pode ser feita mediante cálculos de erros de previsão. Há muitas formas de calcular os erros de previsão, como os estudados por Hyndman e Koehler (2006).

Como os dados analisados nesta tese possuem muitos zeros, as medidas dependentes de escala erro médio (ME), raiz do erro quadrático médio (RMSE) e erro absoluto médio (MAE) serão utilizadas; diferentemente das medidas baseadas em percentuais dos erros, as medidas mencionadas não são impactadas pelas observações iguais a zero.

A vantagem de as medidas baseadas em percentuais dos erros serem independentes de escala não possui relevância no contexto de dados de contagem, uma vez que dados com essa característica possuem a mesma escala.

Considerando e_t o erro de previsão, dado por $e_t = y_t - F_t$, em que y_t é o dado observado e F_t seu valor previsto, as medidas de acurácia ME, RMSE e MAE são definidas por:

$$\text{ME} = \text{média}(e_t),$$

$$\text{RMSE} = \sqrt{\text{média}(e_t^2)} \quad \text{e}$$

$$\text{MAE} = \text{média}(|e_t|).$$

Para qualquer das medidas, quanto menor o valor, menor será o erro de previsão e, portanto, maior acurácia.

CAPÍTULO

4

ESTUDO COM DADOS ARTIFICIAIS

Neste capítulo será realizado um estudo utilizando dados artificiais, para verificar se os modelos ajustados são os mais condizentes com os dados artificiais gerados inicialmente. O apêndice A contém os códigos utilizados para gerar os resultados deste capítulo.

O conjunto de dados artificiais, de tamanho $n = 156$, foi gerado a partir de um modelo Poisson AR(1) ZM, com parâmetros $\phi = 0,8$, $\gamma = -0,7$ e $\delta = 1,3$, sem o uso de variáveis explicativas. Dois modelos foram ajustados: Poisson AR(1) ZM e COM-Poisson AR(1) ZM.

O modelo Poisson AR(1) zero-modificado é escrito como:

$$f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}),$$

com $f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!(1-e^{-\mu_t})}(1-1_{(y_t)})$. Para completar o modelo, as funções de ligação são dadas por:

$$\begin{aligned}\log(\mu_t) &= \phi\log\left(y^{\star}_{t-1}\right)\\ \operatorname{logit}(\omega_t) &= \gamma+\delta\log\left(y^{\star}_{t-1}\right).\end{aligned}$$

No caso do modelo COM-Poisson AR(1) ZM, tem-se:

$$f_{ZMCP}(y_t;\mu_t,\varphi,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTCP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}),$$

em que ZTCP representa a distribuição condicional de COM-Poisson zero truncada, isto é,

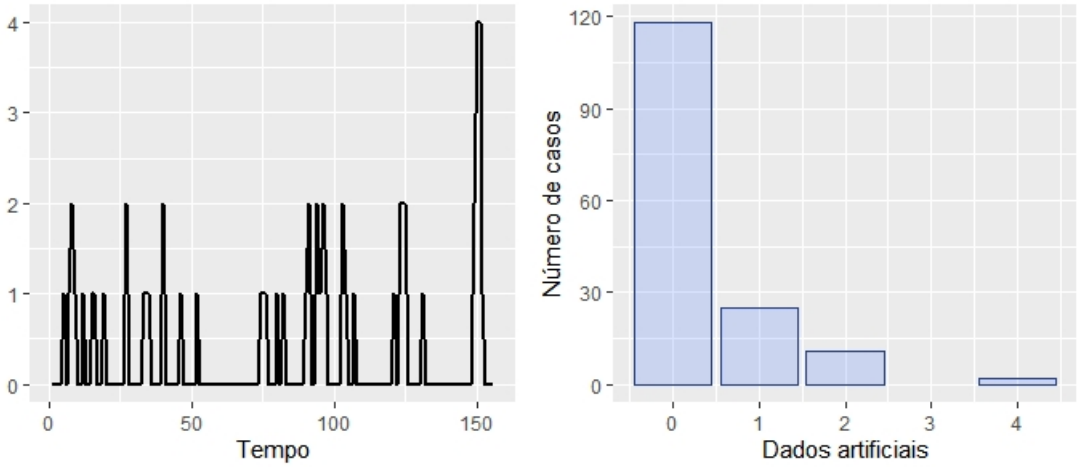
$$f_{ZTCP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) = \frac{1/y_t!^{\varphi}\mu_t^{\varphi y_t}}{\sum\limits_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}-1}(1-1_{(y_t)}).$$

Para completar o modelo, as funções de ligação são dadas por:

$$\log(\mu_t) = \phi \log\left(y^{\star}_{t-1}\right)$$
$$\text{logit}(\omega_t) = \gamma + \delta \log\left(y^{\star}_{t-1}\right).$$

A Figura 4 representa o conjunto de dados gerados.

Figura 4 – Gráfico da série temporal e gráfico de barras referentes aos dados artificiais



4.1 Usando a distribuição Poisson ZM com os dados artificiais

O modelo Poisson ZM foi ajustado aos dados artificiais, utilizando o programa JAGS no R Core Team (2018) por meio dos pacotes "rjags" e "runjags", criados por Plummer (2016) e Denwood (2016), respectivamente.

Nesta aplicação, foram utilizadas 3 cadeias, adaptação (*adapt*) feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1.

Os sumários das distribuições a posteriori do modelo Poisson AR(1) ZM são dados na Tabela 4. A Figura 5 ilustra os intervalos de credibilidade para os parâmetros do modelo. Os intervalos interno e externo contêm 68% e 95% das observações, respectivamente.

As Figuras 6 a 8 apresentam os diagnósticos a posteriori do modelo COM-Poisson AR(1) ZM. O tamanho amostral eficaz (ESS) é dado no quadro superior direito, e no quadro inferior direito é mostrado o erro padrão da estimativa de Monte Carlo (MCSE), calculado pela divisão entre o desvio padrão e a raiz do tamanho amostral eficaz.

Tabela 4 – Sumários das distribuições a posteriori para o modelo Poisson AR(1) ZM para os dados artificiais

Parâmetro	Valor Verdadeiro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	$0,8$	$0,5691$	$0,2540$	$0,0017$	$[0,0643; 1,0542]$
γ	$-0,7$	$-0,5196$	$0,2556$	$0,0025$	$[-1,0283; -0,0240]$
δ	$1,3$	$1,5423$	$0,4090$	$0,0040$	$[0,7506; 2,3530]$

Figura 5 – Intervalos de credibilidade para os parâmetros do modelo Poisson AR(1) ZM para os dados artificiais

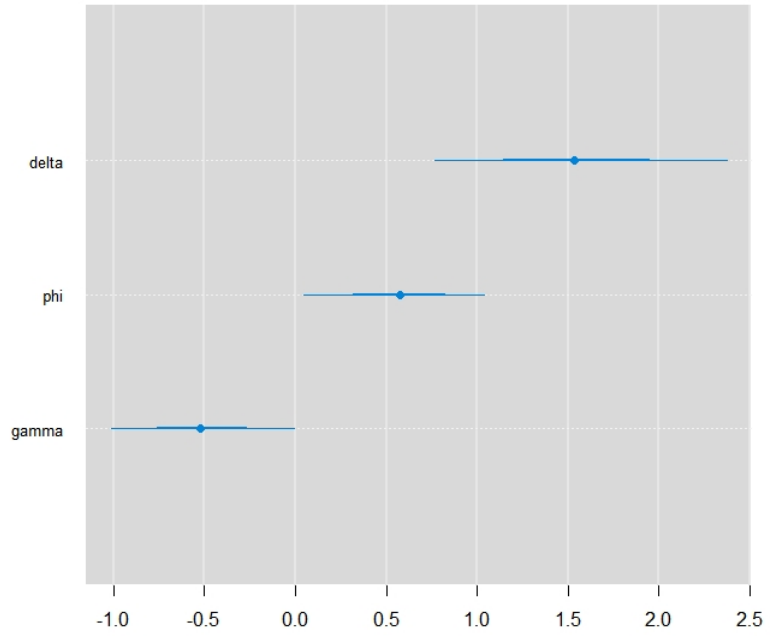


Figura 6 – Diagnóstico MCMC do parâmetro ϕ do modelo Poisson AR(1) ZM para os dados artificiais

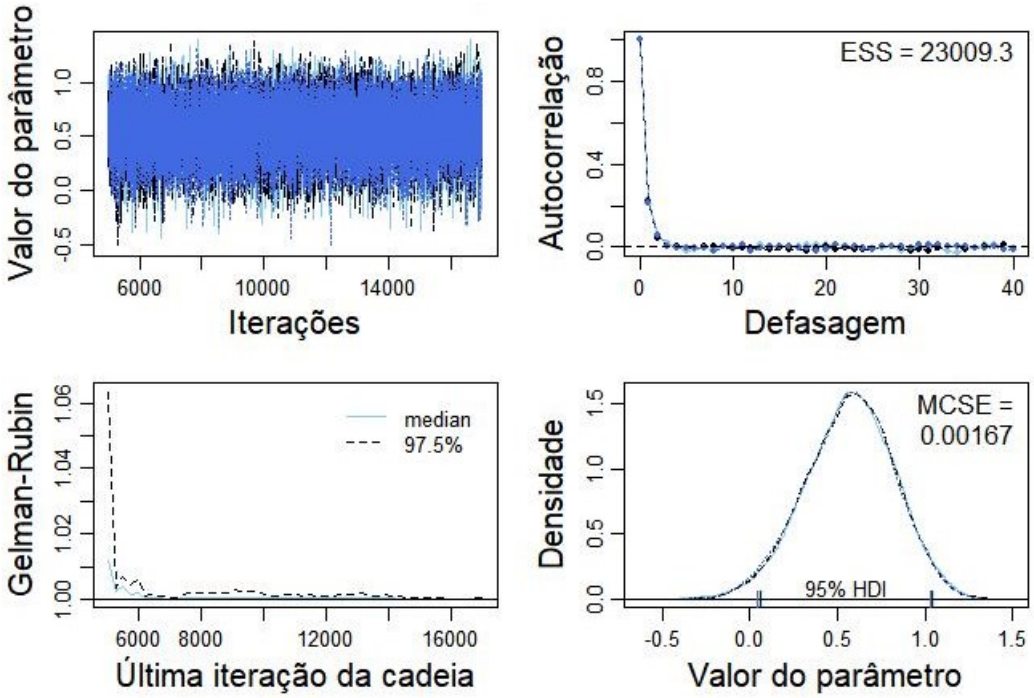


Figura 7 – Diagnóstico MCMC para o parâmetro γ do modelo Poisson AR(1) ZM para os dados artificiais

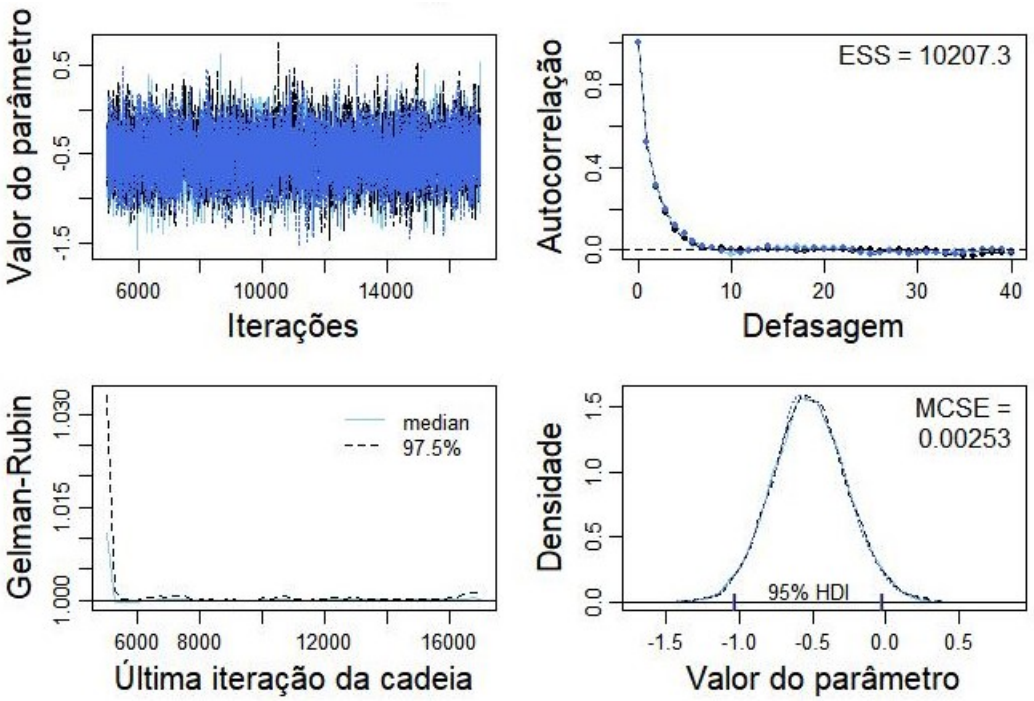
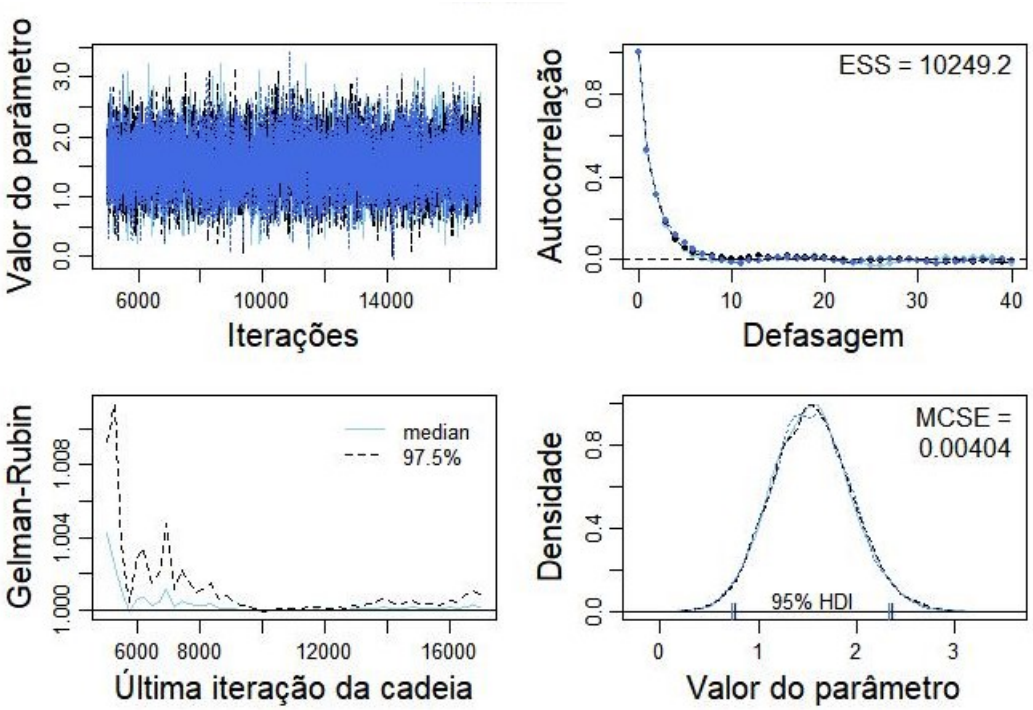


Figura 8 – Diagnóstico MCMC para o parâmetro δ do modelo Poisson AR(1) ZM para os dados artificiais



4.2 Usando a distribuição COM-Poisson ZM com os dados artificiais

O modelo COM-Poisson ZM também foi ajustado aos dados artificiais, utilizando o programa JAGS no R Core Team (2018) por meio dos pacotes "rjags" e "runjags", criados por Plummer (2016) e Denwood (2016), respectivamente.

Nesta aplicação, também foram utilizadas 3 cadeias, adaptação (*adapt*) feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1.

Os sumários das distribuições a posteriori do modelo COM-Poisson AR(1) ZM são dados na Tabela 5. A Figura 9 ilustra os intervalos de credibilidade para os parâmetros do modelo. Os intervalos interno e externo contêm 68% e 95% das observações, respectivamente. As Figuras 10 a 13 apresentam os diagnósticos a posteriori do modelo COM-Poisson AR(1) ZM.

Note que nenhum parâmetro inclui o valor zero dentro dos intervalos de credibilidade; no entanto, o intervalo de credibilidade do parâmetro φ inclui o valor 1, o que mostra que o modelo ajustado não exclui a possibilidade de os dados serem provenientes da distribuição de Poisson zero-modificada.

Tabela 5 – Sumários das distribuições a posteriori para o modelo COM-Poisson AR(1) ZM para os dados artificiais

Parâmetro	Valor Verdadeiro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	$0,8$	$0,5460$	$0,2348$	$0,0016$	$[0,0888; 1,0040]$
γ	$-0,7$	$-0,5249$	$0,2521$	$0,0025$	$[-1,0182; -0,0334]$
δ	$1,3$	$1,5376$	$0,4059$	$0,0040$	$[0,7702; 2,3586]$
φ	$1,0$	$1,3067$	$0,2682$	$0,0019$	$[0,7907; 1,8207]$

Figura 9 – Intervalos de credibilidade para os parâmetros do modelo COM-Poisson AR(1) ZM para os dados artificiais

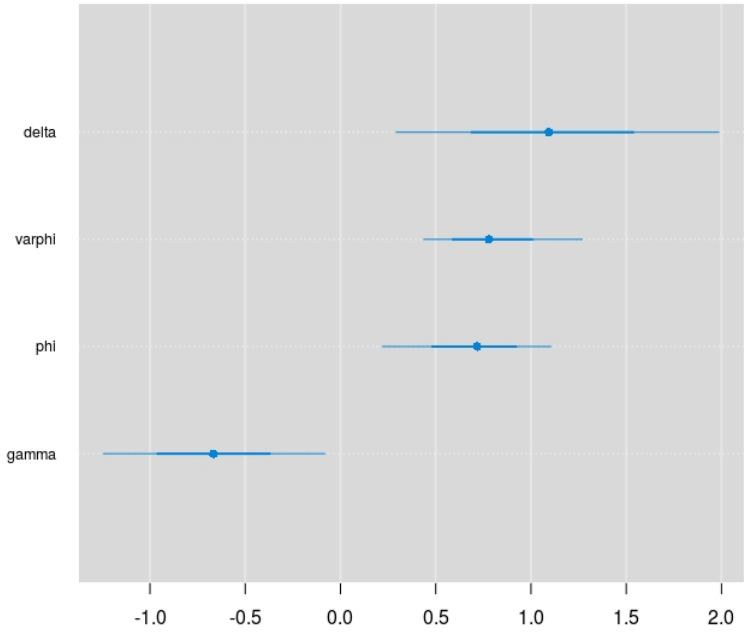


Figura 10 – Diagnóstico MCMC do parâmetro ϕ do modelo COM-Poisson AR(1) ZM para os dados artificiais

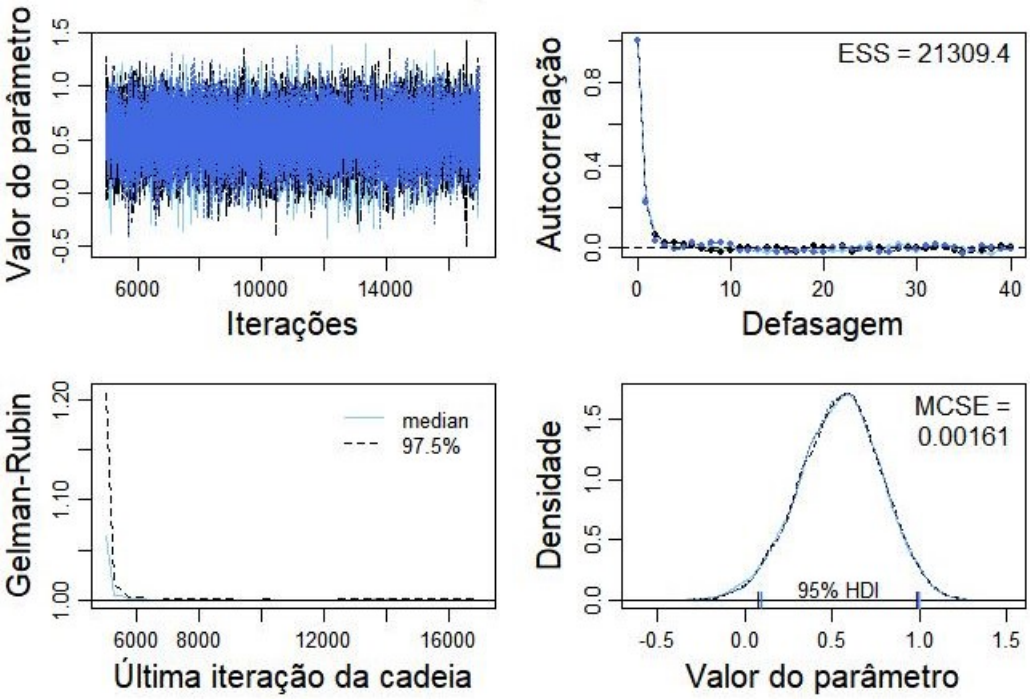


Figura 11 – Diagnóstico MCMC para o parâmetro γ do modelo COM-Poisson AR(1) ZM para os dados artificiais

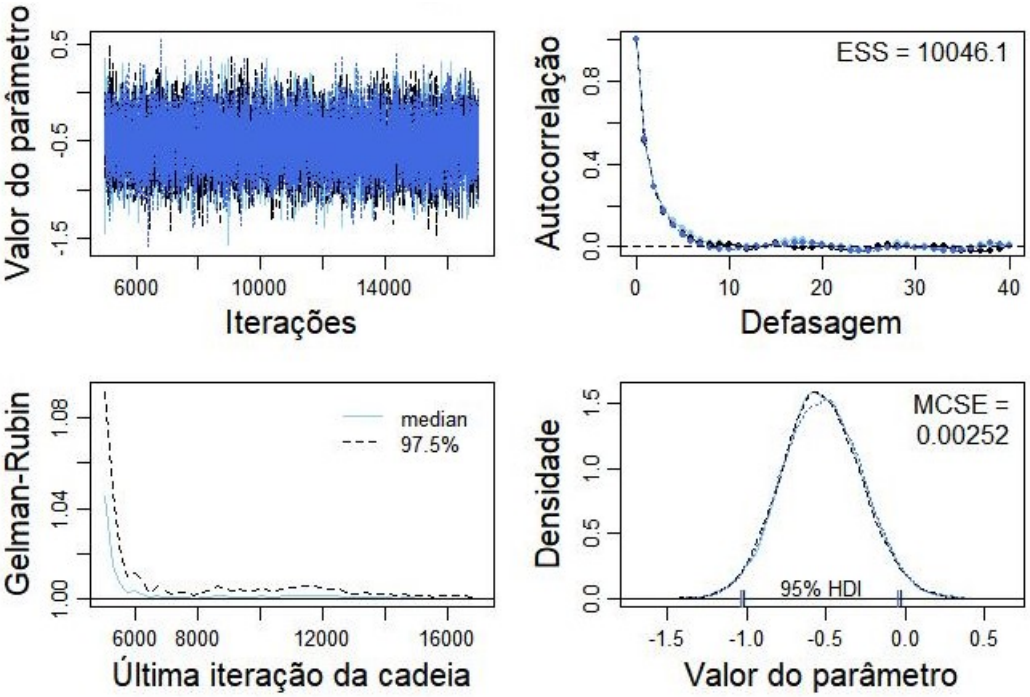


Figura 12 – Diagnóstico MCMC para o parâmetro δ do modelo COM-Poisson AR(1) ZM para os dados artificiais

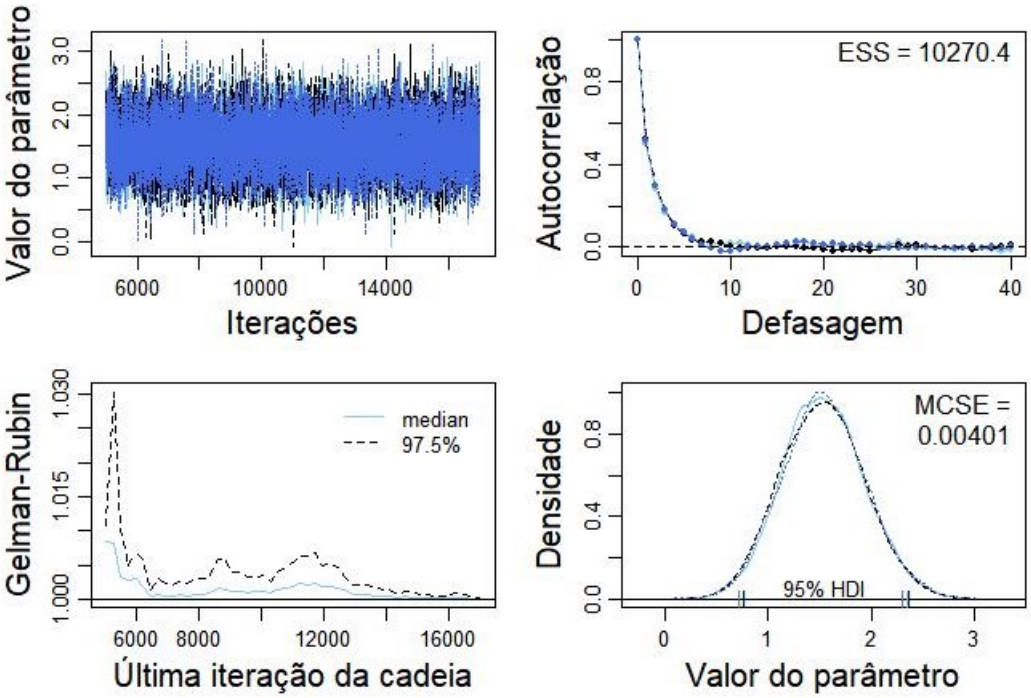
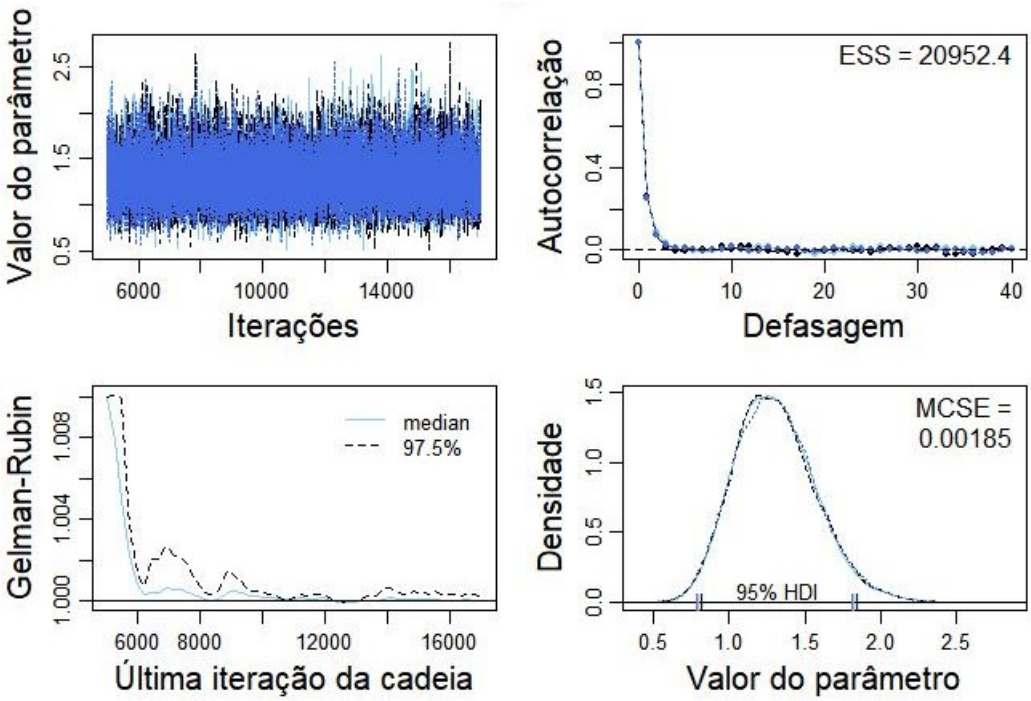


Figura 13 – Diagnóstico MCMC para o parâmetro φ do modelo COM-Poisson AR(1) ZM para os dados artificiais



4.3 Comparação entre os modelos analisados na aplicação com os dados artificiais

Os valores de DIC para os modelos Poisson AR(1) ZM e COM-Poisson AR(1) ZM foram 3080 e 3079, respectivamente. Isso indica que, apesar de haver um parâmetro a mais para ser estimado no modelo COM-Poisson ZM, isso não afetou negativamente o ajuste, uma vez que a distribuição COM-Poisson comtempla a distribuição Poisson como um caso especial e a penalidade devida ao parâmetro adicional não foi significativa.

Os valores de EBIC também foram calculados; no entanto, devido à similaridade com os valores de DIC, assim como não alteração nas conclusões, os valores de EBIC não foram replicados na Tabela 6.

Tabela 6 – Valores de DIC para os modelos Poisson ZM e COM-Poisson ZM ajustados aos dados artificiais

DIC	AR(1)
Poisson ZM	3080
COM-Poisson ZM	3079

CAPÍTULO

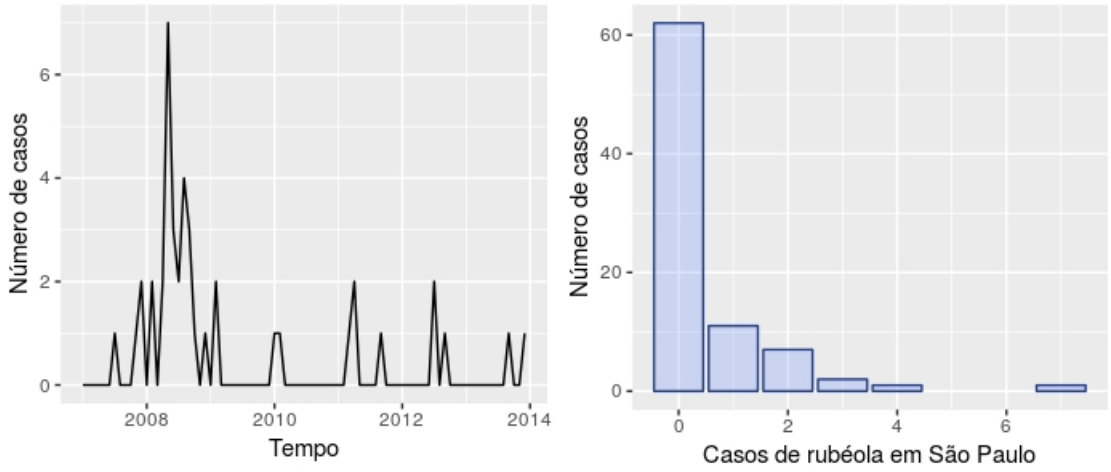
5

APLICAÇÕES EM DADOS REAIS

Neste capítulo serão realizadas aplicações em dois conjuntos de dados reais. O primeiro conjunto de dados analisado para esta tese trata-se de contagens de casos confirmados de síndrome da rubéola congênita (SRC) no estado de São Paulo entre os anos de 2007 e 2013[1].

Os dados são representados na Figura 14. Note a quantidade demasiada de observações iguais a zero no conjunto de dados. Isso se deve ao fato de a rubéola congênita estar bem controlada no Brasil; no entanto é de grande importância o monitoramento da doença devido à possibilidade de haver epidemias.

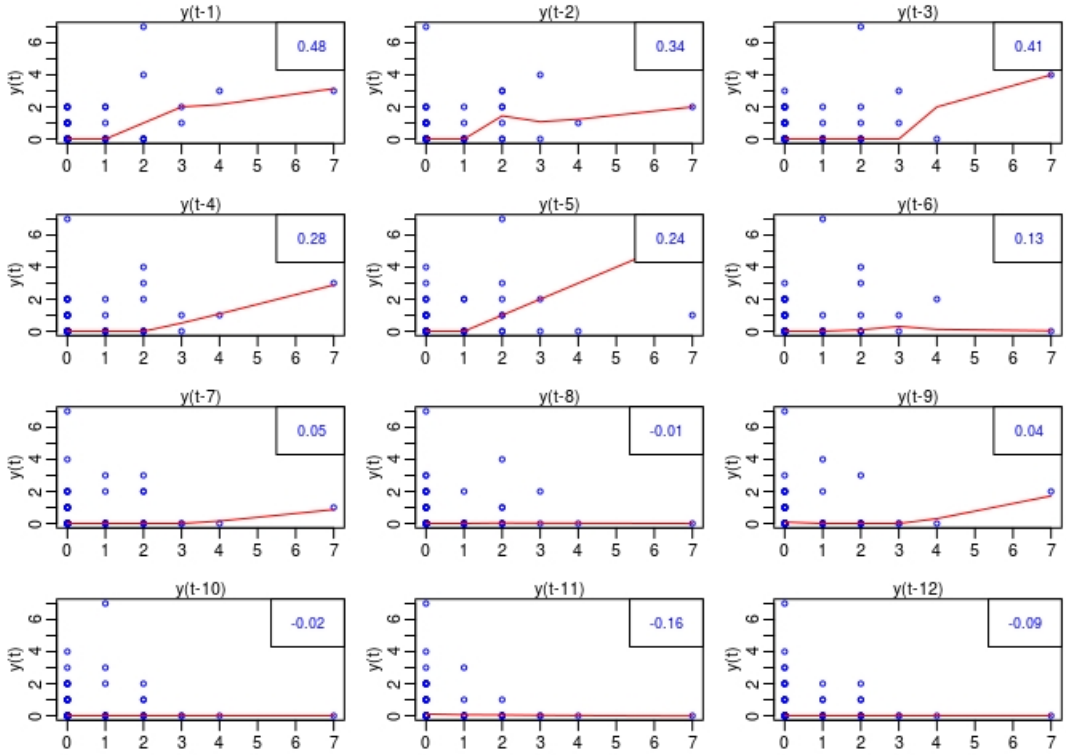
Figura 14 – Gráfico da série temporal e gráfico de barras referentes à SRC no estado de São Paulo



A Figura 15 mostra os gráficos da variável $\boldsymbol{y}'$, representando os dados de rubéola versus suas defasagens. No canto superior direito de cada gráfico há a estimativa de autocorrelação entre as observações no tempo t e $t-i$, $i = 1, \ldots, 12$.

[1] Os dados podem ser recuperados por meio do link <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/srubeolacsp.def>, selecionando como linha o ano do diagnóstico e como coluna o respectivo mês.

Figura 15 – Gráfico de defasagens para os dados de SRC



Por meio da Figura 15, pode-se perceber que de forma geral as autocorrelações diminuem quanto maior a defasagem e a autocorrelação mais forte ocorre quando a defasagem é igual a 1.

5.1 Usando a distribuição Poisson ZM com os dados de SRC

Para ilustrar a primeira aplicação, foi escolhido primeiramente o modelo mais simples, isto é, autorregressivo de primeira ordem. Assim, fazendo uso da equação (2.5), o modelo Poisson AR(1) ZM é escrito como:

$$f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}),$$

com $f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!(1-e^{-\mu_t})}(1-1_{(y_t)})$. Para completar o modelo, as funções de ligação são dadas por:

$$\log(\mu_t) = \beta + \phi\log\left(y^{\star}_{t-1}\right)$$
$$\text{logit}(\omega_t) = \gamma + \delta\log\left(y^{\star}_{t-1}\right).$$

O modelo Poisson AR(1) ZM foi aplicado para ajustar os dados de SRC. O programa JAGS foi utilizado por meio dos pacotes "rjags" e "runjags" no R Core Team (2018), criados por Plummer (2016) e Denwood (2016), respectivamente.

O modelo com o parâmetro β resultou em parâmetros não significativos; assim, optou-se pelo modelo Poisson AR(1) sem o parâmetro β. Os sumários a posteriori para os parâmetros são disponibilizados na Tabela 7 e os resultados de diagnóstico são apresentados nas Figuras 16 a 18.

Tabela 7 – Sumários das distribuições a posteriori para o modelo Poisson AR(1) ZM para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	0,7180	0,1911	0,00129	[0,3379; 1,0836]
γ	−0,6673	0,2993	0,00248	[−1,2550; −0,0749]
δ	1,1085	0,4284	0,00351	[0,2830; 1,9588]

Figura 16 – Diagnóstico MCMC para o parâmetro ϕ do modelo Poisson AR(1) ZM para os dados de SRC

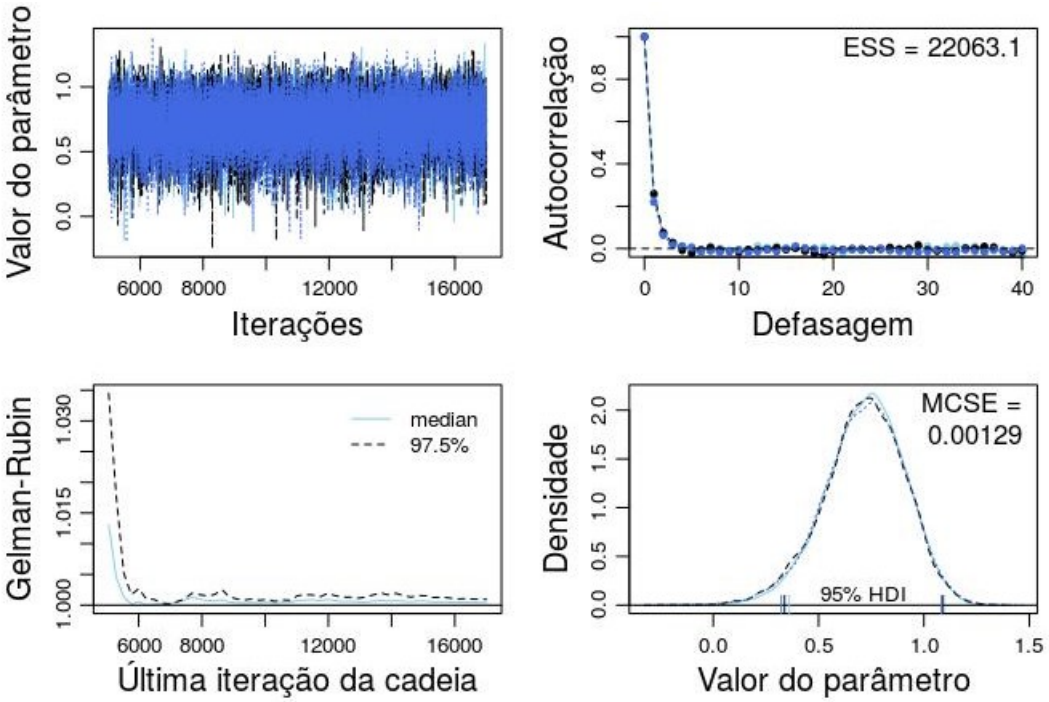


Figura 17 – Diagnóstico MCMC para o parâmetro γ do modelo Poisson AR(1) ZM para os dados de SRC

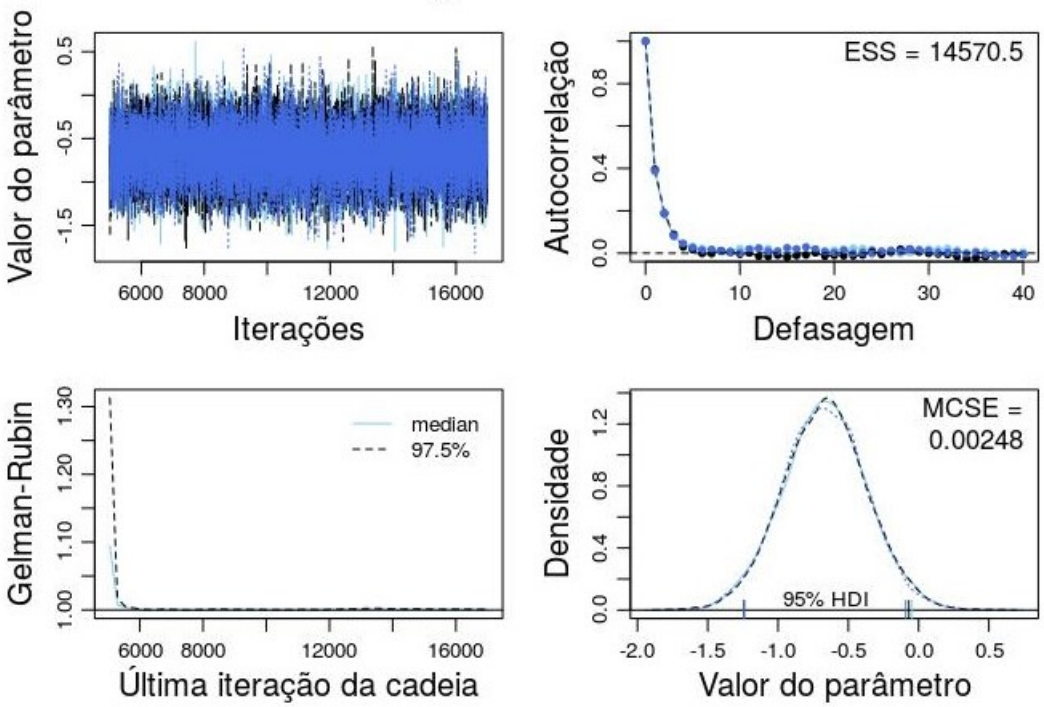
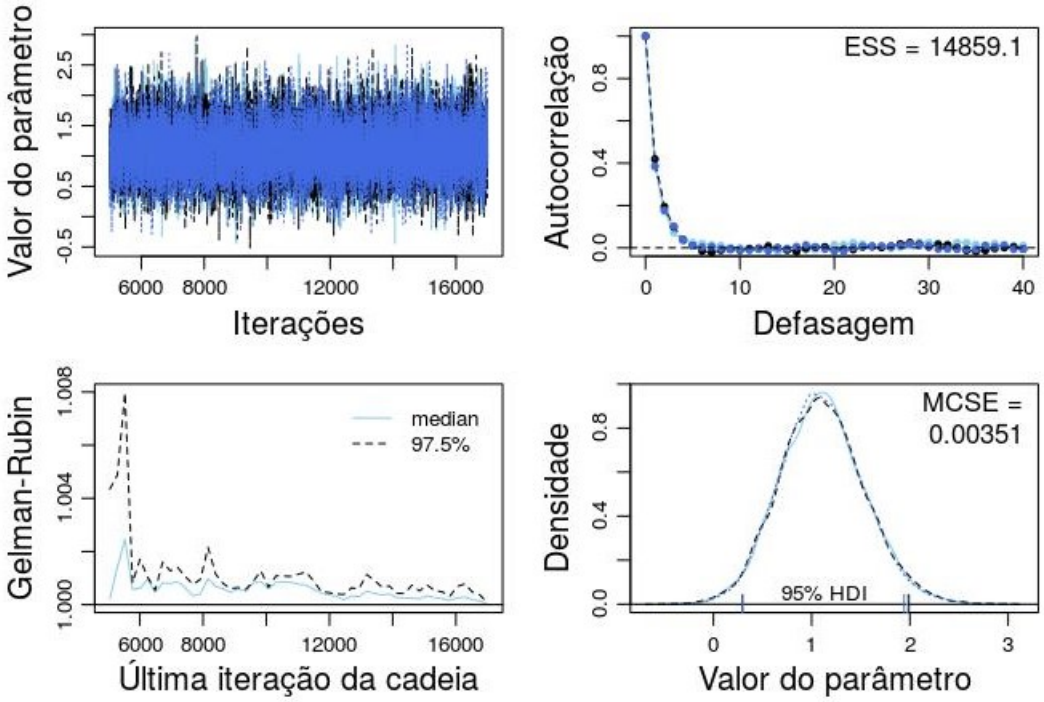


Figura 18 – Diagnóstico MCMC para o parâmetro δ do modelo Poisson AR(1) ZM para os dados de SRC



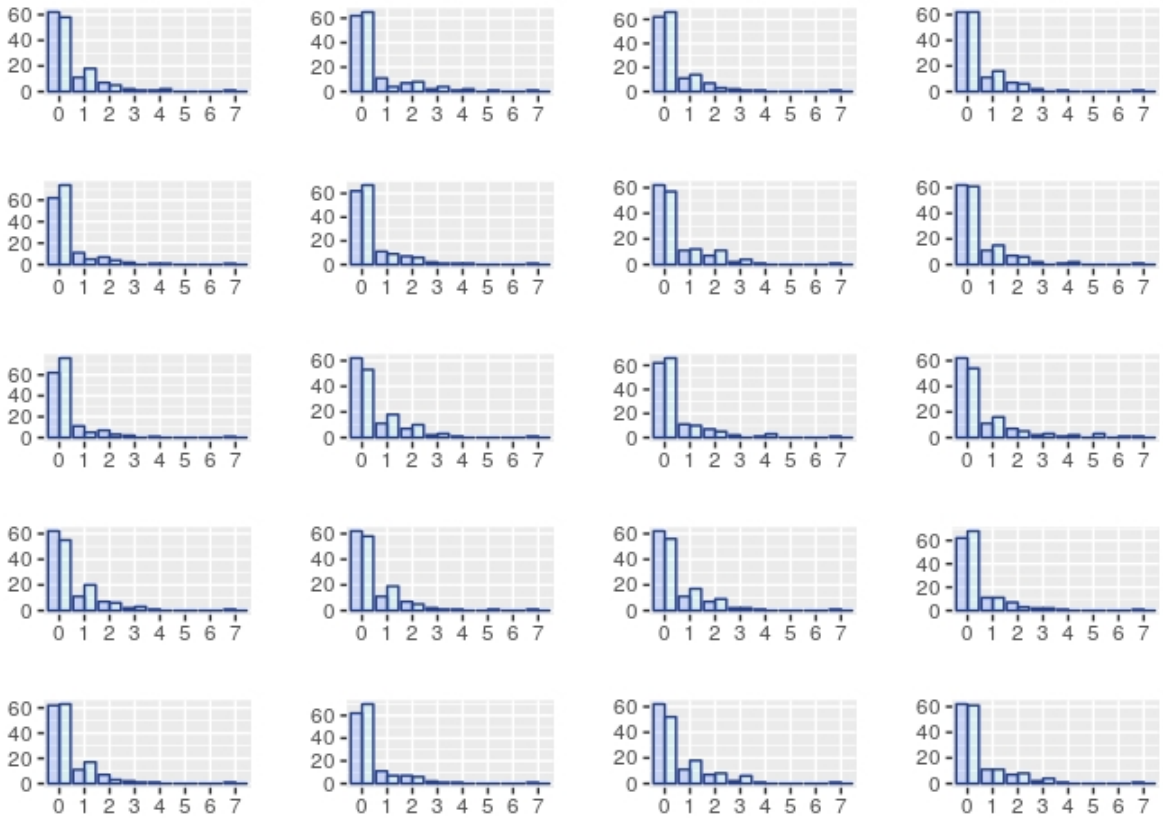
Para realizar os gráficos de diagnóstico e calcular os intervalos HDI também foi utilizado o pacote "DBDA2E-utilities", criado por Kruschke (2015). Nesta aplicação foram utilizadas 3 cadeias, adaptação (*adapt*) feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1.

Note no canto superior direito das Figuras 16, 17 e 18 o tamanho amostral eficaz, que foi acima de 14000 para todos os parâmetros.

Os quadros à esquerda nas Figuras 16, 17 e 18 indicam convergência das cadeias. As estimativas das curvas das distribuições a posteriori são mostradas no canto inferior direito. Note como as curvas se sobrepõem, indicando uma vez mais a convergência das cadeias.

A Figura 19 mostra o resultado do procedimento de checagem preditiva a posteriori. Foram gerados 20 conjuntos de dados fictícios do modelo Poisson AR(1) ZM com 84 observações em cada conjunto. Para gerar as réplicas, foram amostrados 20 valores de cada um dos parâmetros estimados pelo método MCMC, isto é, $\hat{\phi}$, $\hat{\gamma}$ e $\hat{\delta}$.

Figura 19 – Comparação entre valores observados e ajustados para o modelo Poisson AR(1) ZM para os dados de SRC. As barras à esquerda representam os dados observados e à direita os dados ajustados



Pode-se perceber que, de maneira geral, o modelo consegue captar o excesso de zeros; no entanto, para essas 20 réplicas, o modelo não gerou observações superiores a 6, o que ocorre nos dados de rubéola (aparecem 7 casos no mês de maio de 2008, por exemplo). Talvez ordens superiores ou outra distribuição base, como a COM-Poisson, sejam capazes de gerar observações maiores.

Para testar novamente o uso da distribuição Poisson ZM para ajustar os dados de SRC,

os modelos Poisson AR(2) ZM e Poisson ARMA(1,1) ZM também foram implementados. O modelo Poisson AR(2) é escrito da seguinte maneira:

$$f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}).$$

Para completar o modelo, as funções de ligação são dadas por:

$$\begin{aligned}\log(\mu_t) &= \beta + \phi_1 \log\left(y^{\star}_{t-1}\right) + \phi_2 \log\left(y^{\star}_{t-2}\right)\\ \text{logit}(\omega_t) &= \gamma + \delta_1 \log\left(y^{\star}_{t-1}\right) + \delta_2 \log\left(y^{\star}_{t-2}\right).\end{aligned}$$

Os modelos com os parâmetros β e/ou γ resultou em parâmetros não significativos, assim, o modelo sem estes dois parâmetros foi analisado.

Para esse modelo, foi utilizado $thin = 5$, para diminuir a autocorrelação entre os elementos das cadeias e, assim, aumentar o tamanho amostral eficaz. A Tabela 8 mostra os sumários das distribuições a posteriori dos parâmetros do modelo Poisson AR(2) ZM.

Tabela 8 – Sumários das distribuições a posteriori para o modelo Poisson AR(2) ZM para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ_1	0,9291	0,2488	0,00143	[0,4298; 1,3995]
ϕ_2	−0,4321	0,3362	0,00191	[−1,0980; 0,2183]
δ_1	0,8530	0,4689	0,00252	[−0,0501; 1,7801]
δ_2	1,4013	0,4806	0,00255	[0,4574; 2,3451]

As Figuras 20 a 23 mostram o diagnóstico das cadeias geradas. Note que o tamanho amostral eficaz foi superior a 30000 para todos os parâmetros do modelo.

Figura 20 – Diagnóstico MCMC para o parâmetro ϕ_1 do modelo Poisson AR(2) ZM para os dados de SRC

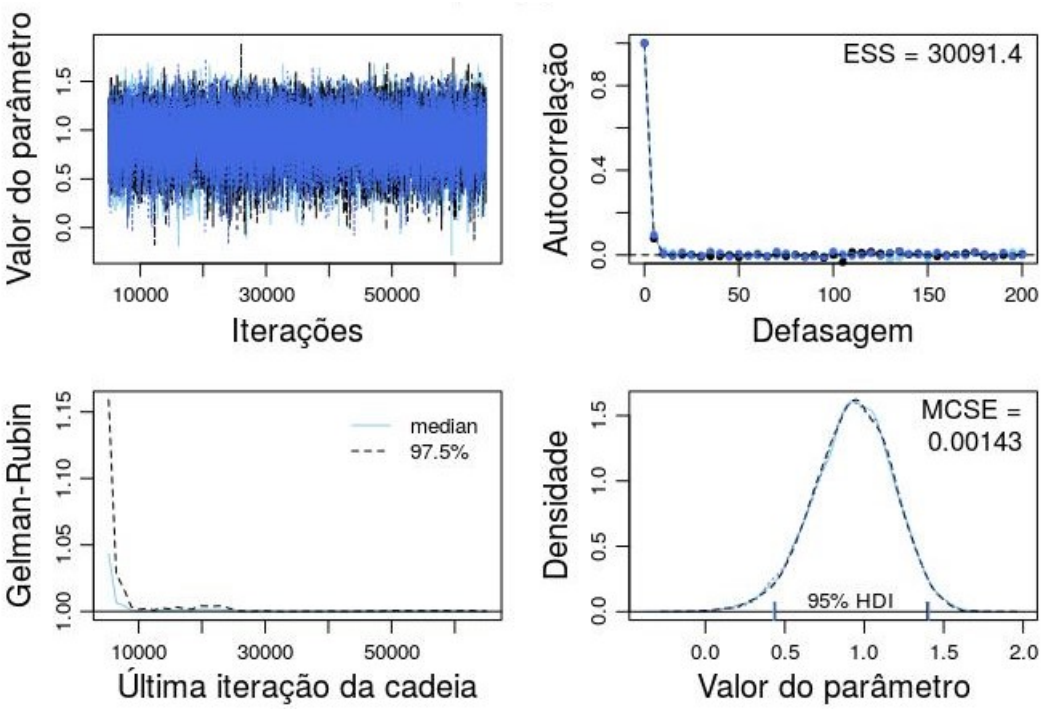


Figura 21 – Diagnóstico MCMC para o parâmetro ϕ_2 do modelo Poisson AR(2) ZM para os dados de SRC

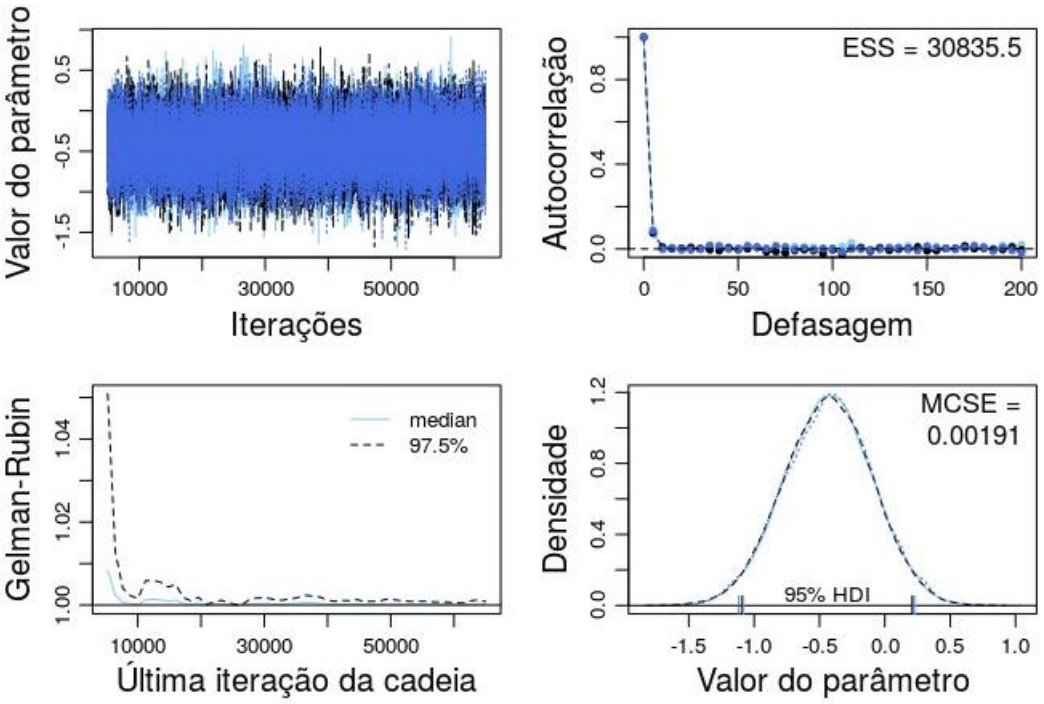


Figura 22 – Diagnóstico MCMC para o parâmetro δ_1 do modelo Poisson AR(2) ZM para os dados de SRC

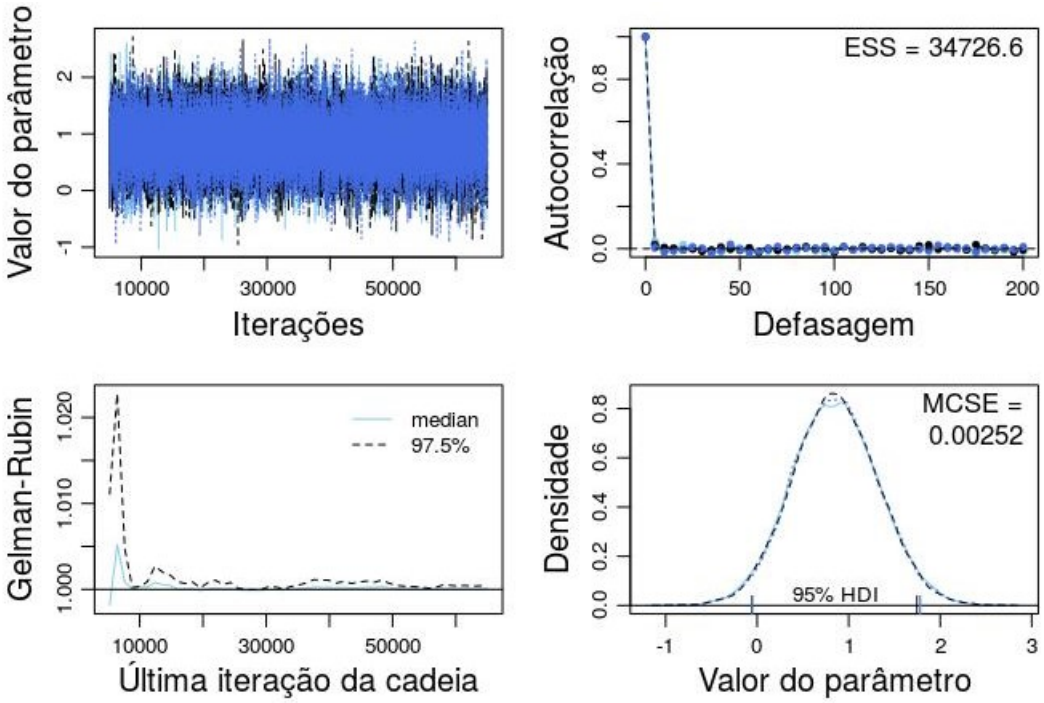


Figura 23 – Diagnóstico MCMC para o parâmetro δ_2 do modelo Poisson AR(2) ZM para os dados de SRC

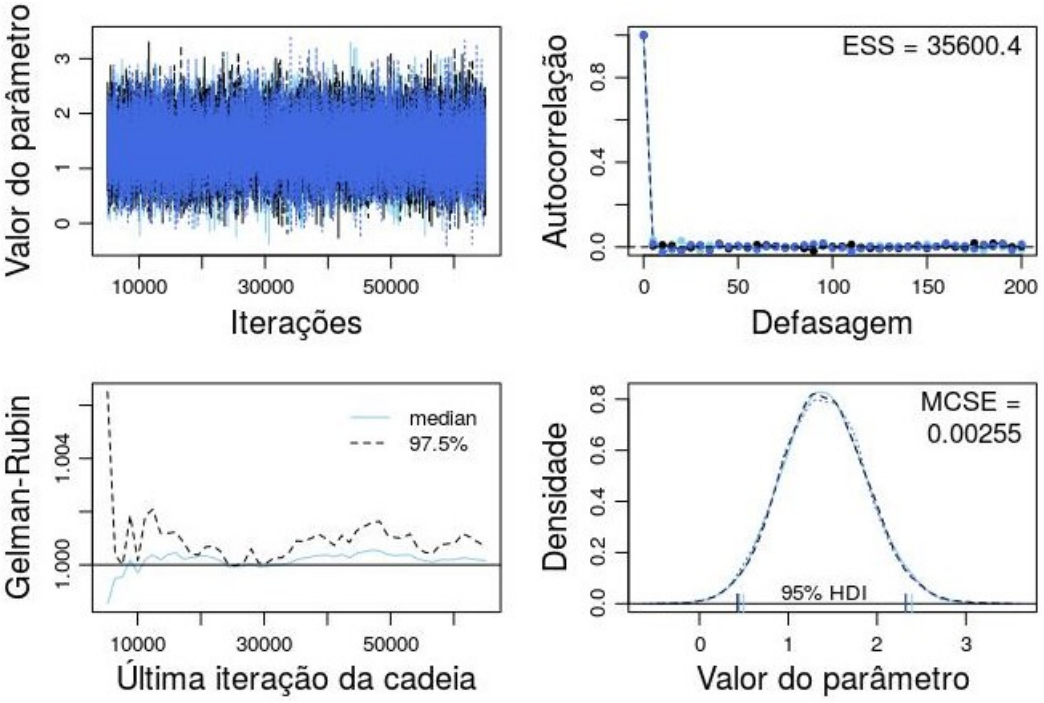
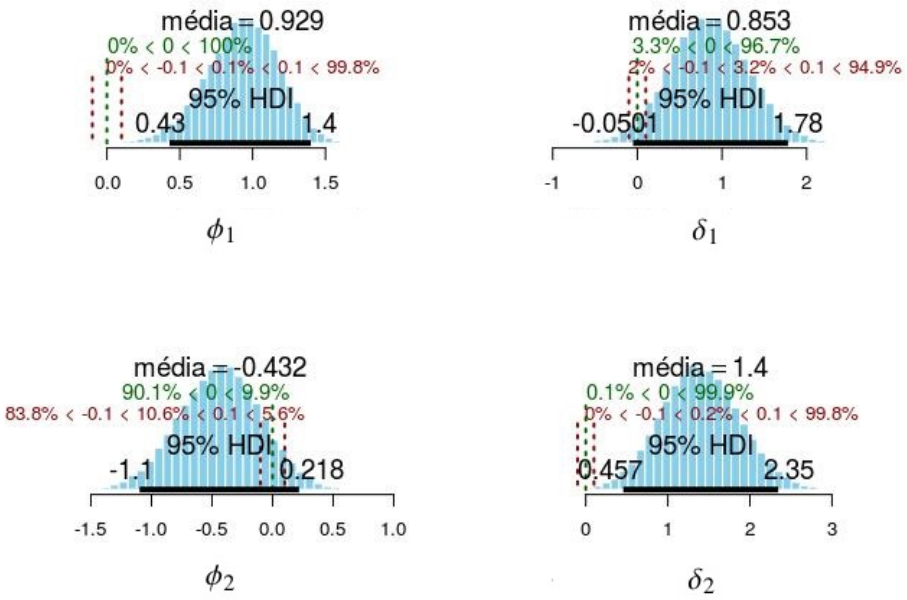


Figura 24 – Distribuições a posteriori do modelo Poisson AR(2) ZM com regiões de equivalência prática para os dados de SRC



Analisando os quadros à esquerda das Figuras 20 a 23 verifica-se que as cadeias parecem ter convergido, pelo aspecto visual (quadros superiores à esquerda) e pelo critério numérico (quadros inferiores à esquerda), por meio da estatística de Gelman-Rubin.

Por meio dos intervalos de credibilidade (HDI (95%)) dispostos na Tabela 8, pode-se perceber que o valor zero está incluído em dois intervalos de credibilidade: os referentes aos parâmetros ϕ_2 e δ_1. Uma análise mais detalhada é mostrada na Figura 24.

Note que os HDIs dos parâmetros ϕ_1 e δ_2 excluem a região de equivalência prática (ROPE) em torno do valor zero; no entanto, isso não ocorre nos HDIs dos parâmetros ϕ_2 e δ_2. No caso do parâmetro ϕ_2, aproximadamente 11% das observações geradas desse parâmetro estão na ROPE que abrange os valores de $-0,1$ a $0,1$, enquanto que aproximadamente 3% das observações geradas do parâmetro δ_1 estão no intervalo de $-0,1$ a $0,1$. Assim, os HDIs dos parâmetros ϕ_2 e δ_1 não excluem a ROPE em torno de zero. Portanto, o modelo Poisson AR(2) ZM não se mostrou melhor que o modelo mais simples, Poisson AR(1) ZM, uma vez que dois de seus parâmetros não são significativos.

O modelo Poisson ARMA(1,1) também foi implementado, usando *thin*=5. A equação do modelo Poisson ARMA(1,1) ZM é como segue:

$$
\begin{aligned}
f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) &= (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}) \\
\log(\mu_t) &= \phi\log\left(y^{\star}_{t-1}\right) + \theta\log\left(y^{\star}_{t-1}/\mu_{t-1}\right) \\
\text{logit}(\omega_t) &= \delta\log\left(y^{\star}_{t-1}\right).
\end{aligned}
$$

A Tabela 9 mostra os resultados a posteriori e os diagnósticos a posteriori são apresentados nas Figuras 25 a 27.

Tabela 9 – Sumários das distribuições a posteriori para o modelo Poisson ARMA(1,1) ZM para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	0,4315	0,2981	0,00189	[−0,1719; 0,9884]
θ	0,5401	0,2096	0,00147	[0,1268; 0,9033]
δ	1,6189	0,3878	0,00202	[0,8723; 2,3891]

Figura 25 – Diagnóstico MCMC para o parâmetro ϕ do modelo Poisson ARMA(1,1) ZM para os dados de SRC

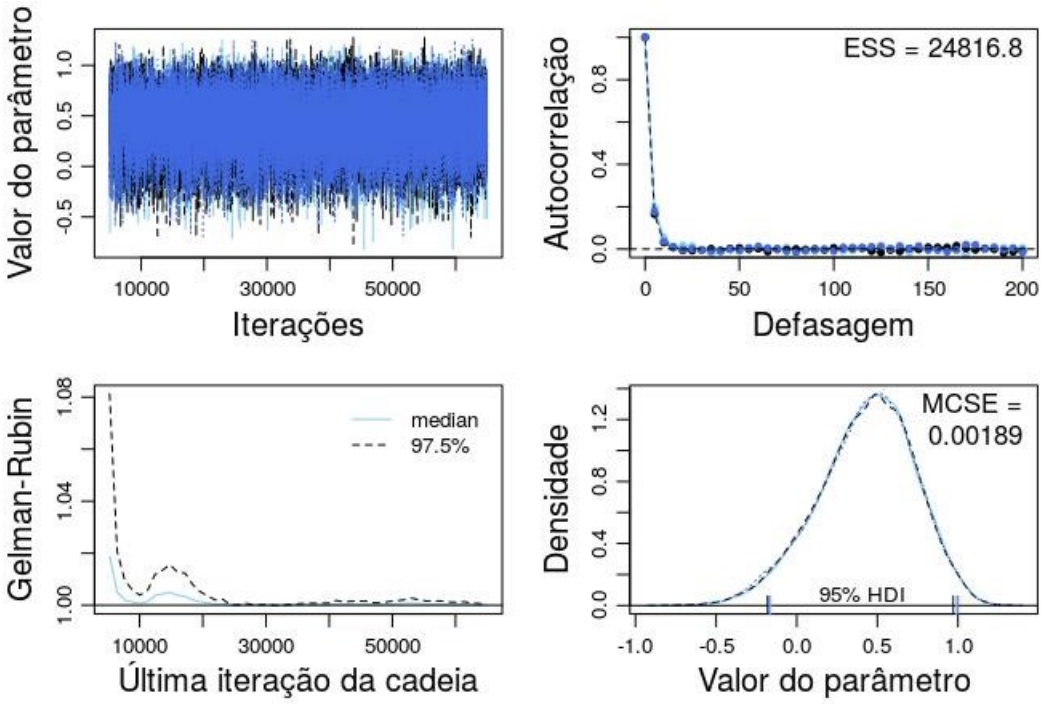


Figura 26 – Diagnóstico MCMC para o parâmetro θ do modelo Poisson ARMA(1,1) ZM para os dados de SRC

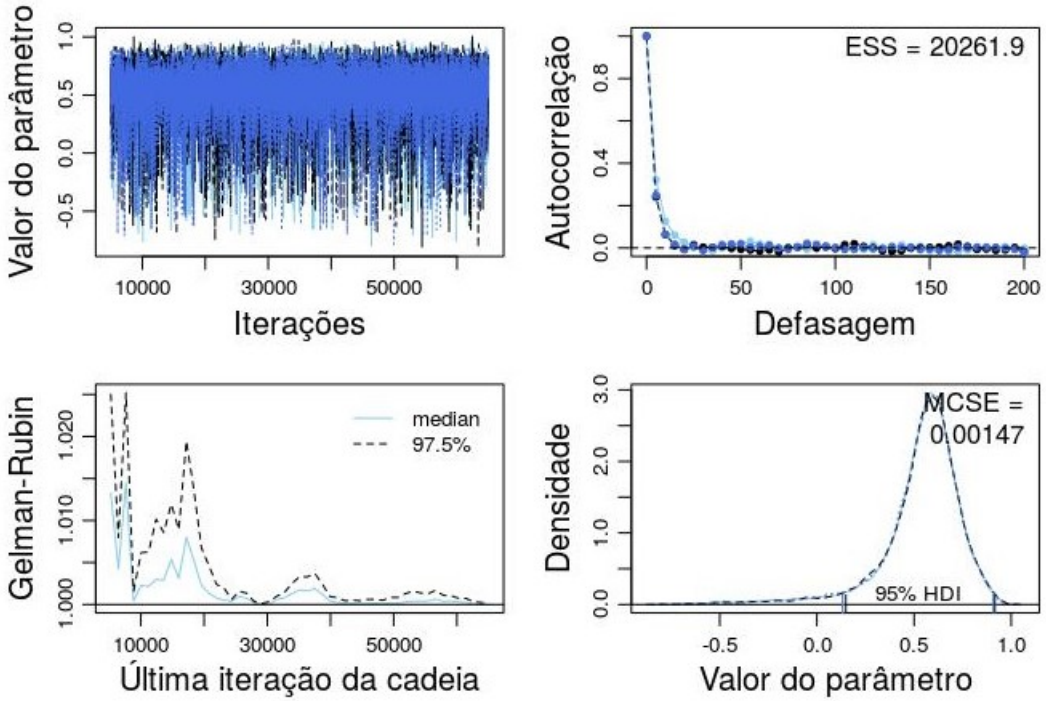
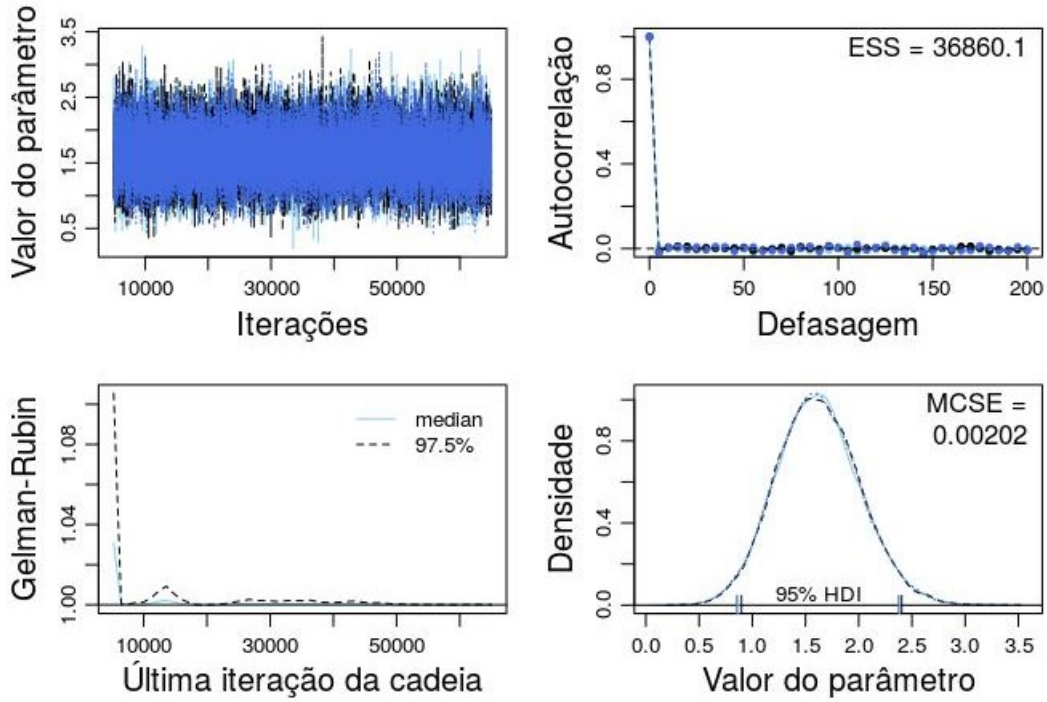


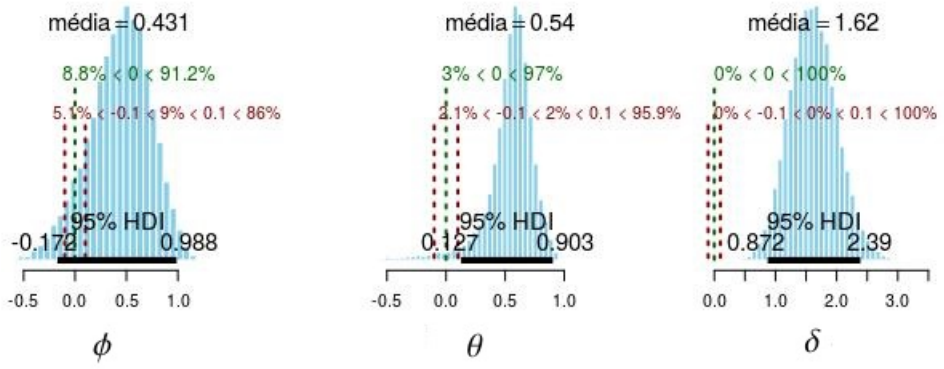
Figura 27 – Diagnóstico MCMC para o parâmetro δ do modelo Poisson ARMA(1,1) ZM para os dados de SRC



A Figura 28 mostra as ROPEs para os parâmetros do modelo Poisson ARMA(1,1) ZM. Observe como o HDI do parâmetro ϕ inclui a ROPE que abrange o intervalo de $-0,1$ a $0,1$. A ROPE que abrange o mesmo intervalo quase não é excluída pelo HDI do parâmetro θ. Assim,

entre os modelos ajustados com base na distribuição de Poisson ZM, o de menor ordem foi capaz de explicar melhor os dados e cujos parâmetros são todos significativos.

Figura 28 – Distribuições a posteriori do modelo Poisson ARMA(1,1) ZM com regiões de equivalência prática para os dados de SRC



Antes de passar para os próximos modelos, é importante notar a importância da checagem preditiva a posteriori. Como forma de ilustração dessa afirmação, o modelo Poisson AR(1), sem modificação no zero, foi utilizado para ajustar os dados de rubéola. Para esse modelo não foi necessário utilizar *thin* diferente de 1. A equação do modelo Poisson AR(1) é como segue:

$$f_P(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!}.$$

Para completar o modelo, a função de ligação é dada por:

$$\log(\mu_t) = \phi\log\left(y_{t-1}^{\star}\right)$$

A Tabela 10 mostra os resultados a posteriori para o parâmetro ϕ do modelo Poisson AR(1).

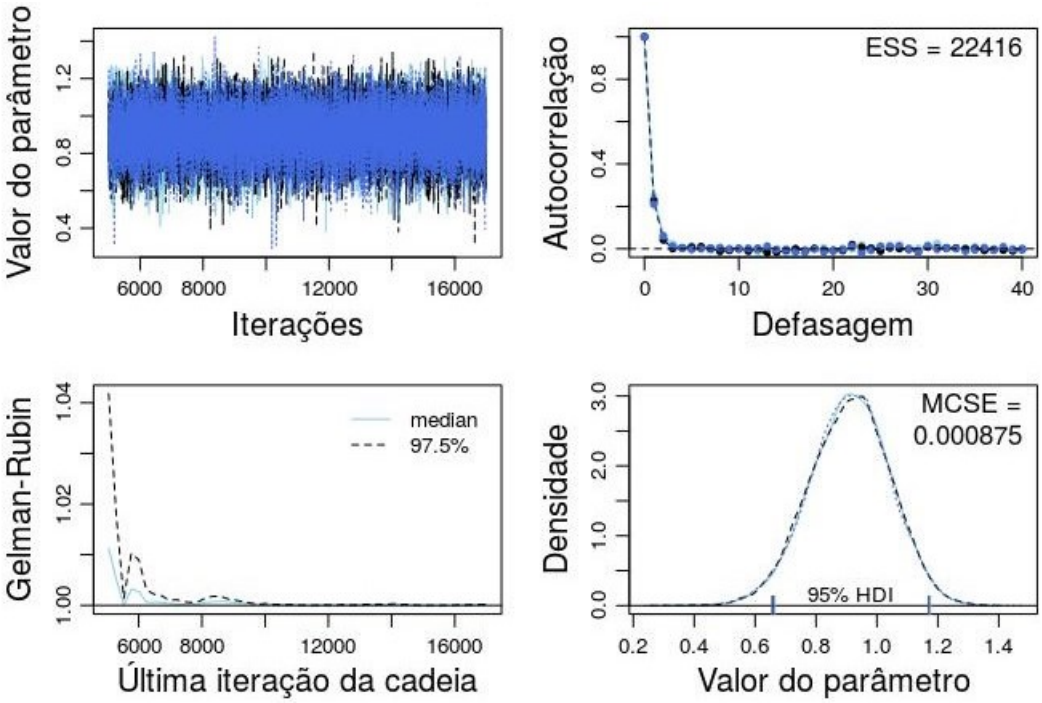
Tabela 10 – Sumários da distribuição a posteriori para o modelo Poisson AR(1) para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	0,9118	0,1298	0,00088	[0,6547; 1,1617]

A Figura 29 mostra o diagnóstico MCMC para o parâmetro ϕ do modelo Poisson AR(1). Note no diagnóstico das cadeias como o tamanho amostral eficaz foi alto (acima de 22000), assim como a Estatística de Gelman-Rubin foi satisfatória.

No entanto, a Figura 30 mostra que esse modelo não conseguiu ajustar bem os dados. Isso ocorre porque o diagnóstico analisa os parâmetros do modelo em si, e não o ajuste do modelo.

Figura 29 – Diagnóstico MCMC para o parâmetro ϕ do modelo Poisson AR(1)



Por isso, é importante a fase de checagem a posteriori para verificar se o modelo escolhido consegue imitar os dados. Aqui cabe ressaltar que a coluna representada pelo número 8 na Figura 30 representa os valores iguais ou superiores a 8.

Por fim, conclui-se que entre os modelos ajustados com base na distribuição de Poisson ZM, o de menor ordem é o que consegue explicar melhor os dados e cujos parâmetros são todos significativos.

5.2 Usando a distribuição COM-Poisson ZM com os dados de SRC

Aplicando a distribuição COM-Poisson, o modelo COM-Poisson ARMA ZM é escrito como

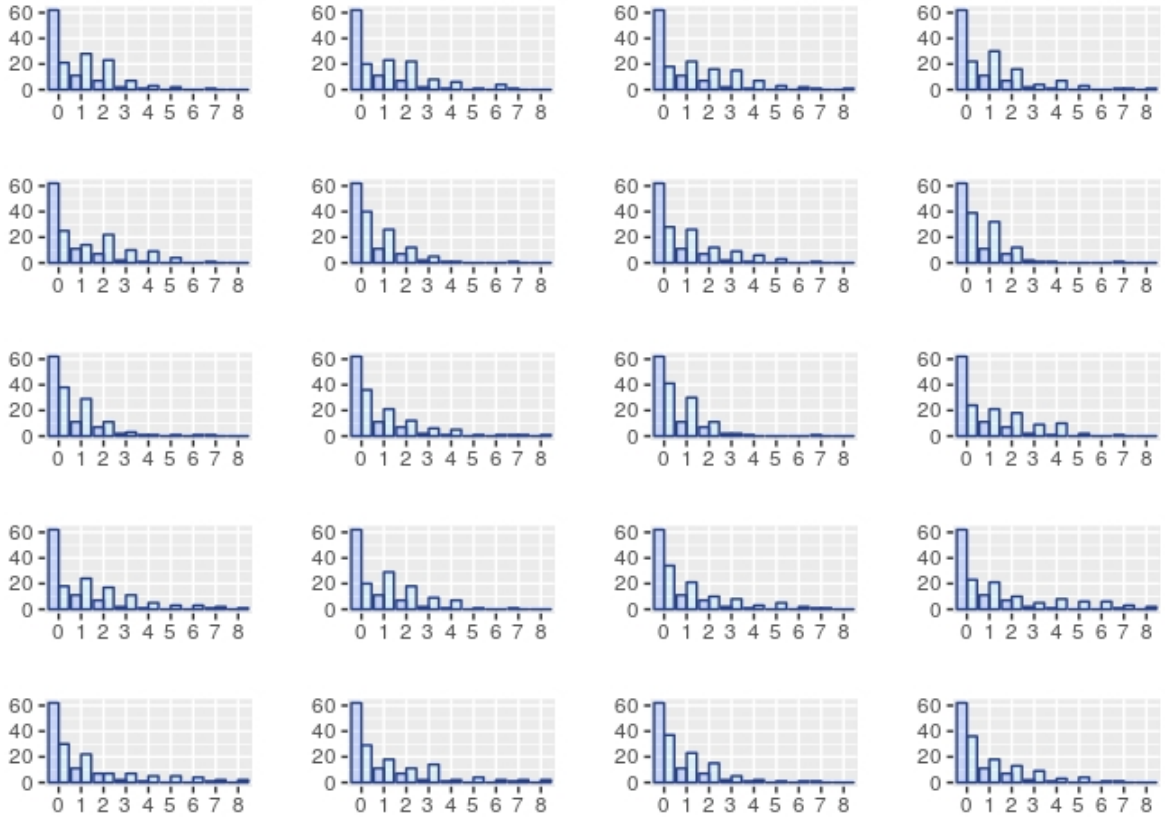
$$f_{ZMCP}(y_t;\mu_t,\varphi,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTCP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}),$$

em que ZTCP representa a distribuição condicional de COM-Poisson zero truncada, isto é,

$$f_{ZTCP}(y_t;\mu_t,\varphi|\mathcal{F}_{t-1}) = \frac{1/y_t!^{\varphi}\mu_t^{\varphi y_t}}{\sum_{n=0}^{\infty}\left(\frac{\mu_t^n}{n!}\right)^{\varphi}-1}(1-1_{(y_t)}).$$

Para completar o modelo, as funções de ligação para o modelo mais simples (COM-Poisson

Figura 30 – Comparação entre valores observados e ajustados para o modelo Poisson AR(1) usando os dados de SRC. As barras à esquerda representam os dados observados e à direita os dados ajustados



AR(1) ZM) são dadas por:

$$\log(\mu_t) = \beta + \phi \log\left(y_{t-1}^{\star}\right)$$
$$\operatorname{logit}(\omega_t) = \gamma + \delta \log\left(y_{t-1}^{\star}\right).$$

O modelo com os parâmetros β e γ resultou em parâmetros não significativos, assim como o modelo com o parâmetro β e sem o parâmetro γ. O modelo com o parâmetro γ e sem o parâmetro β resultou em todos os parâmetros serem significativos.

Nesta aplicação foram utilizadas 3 cadeias, adaptação feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1. Os sumários das distribuições a posteriori do modelo COM-Poisson AR(1) ZM são dados na Tabela 11. O intervalo de credibilidade para o parâmetro φ inclui o valor 1, não excluindo, assim, a possibilidade de os dados terem como base a distribuição Poisson ZM.

A Figura 31 ilustra os intervalos de credibilidade para os parâmetros do modelo. Os intervalos interno e externo contêm 68% e 95% das observações, respectivamente. As Figuras 32 a 35 apresentam os diagnósticos a posteriori do modelo COM-Poisson AR(1) ZM.

Tabela 11 – Sumários das distribuições a posteriori para o modelo COM-Poisson AR(1) ZM para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	$0,7022$	$0,2256$	$0,0016$	$[0,2454; 1,1203]$
γ	$-0,6667$	$0,2973$	$0,0025$	$[-1,2543; -0,0913]$
δ	$1,1089$	$0,4300$	$0,0037$	$[0,2799; 1,9694]$
φ	$0,7986$	$0,2126$	$0,0015$	$[0,4044; 1,2175]$

Figura 31 – Intervalos de credibilidade para os parâmetros do modelo COM-Poisson AR(1) ZM para os dados de SRC

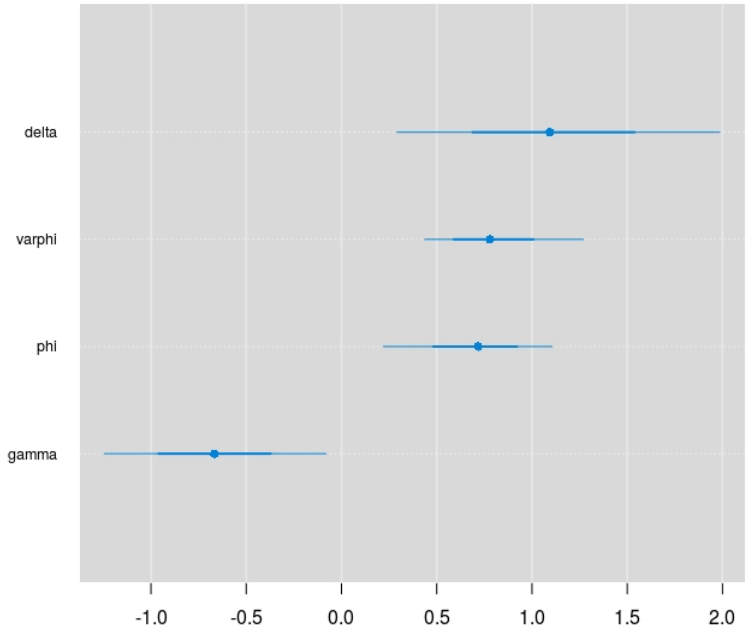


Figura 32 – Diagnóstico MCMC do parâmetro ϕ do modelo COM-Poisson AR(1) ZM para os dados de SRC

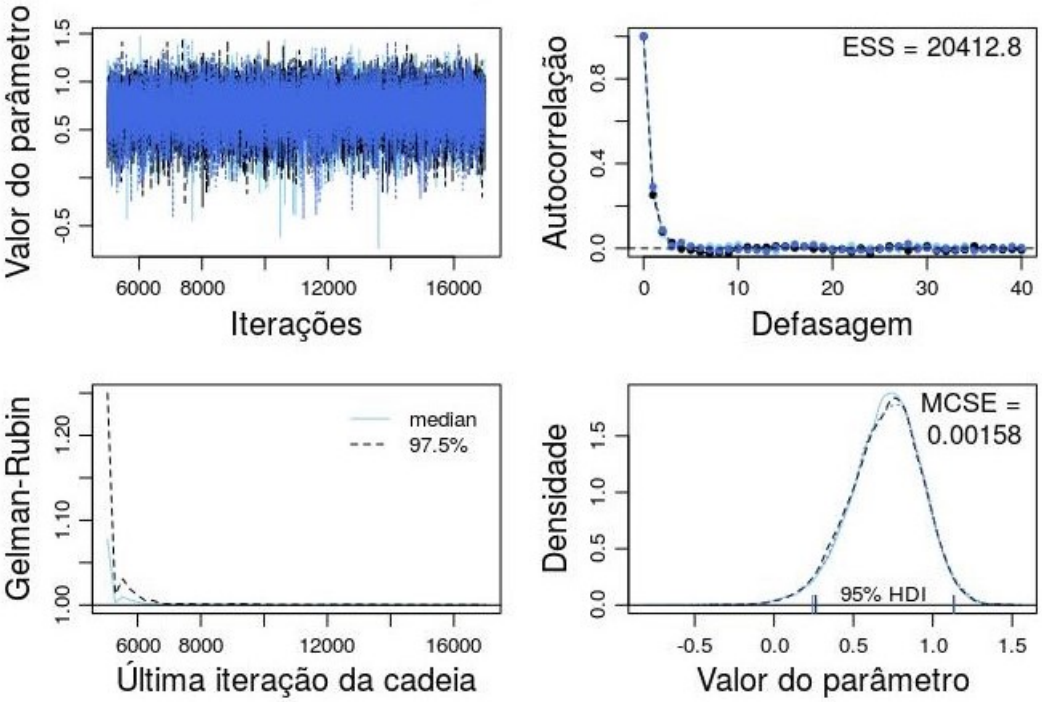


Figura 33 – Diagnóstico MCMC para o parâmetro γ do modelo COM-Poisson AR(1) ZM para os dados de SRC

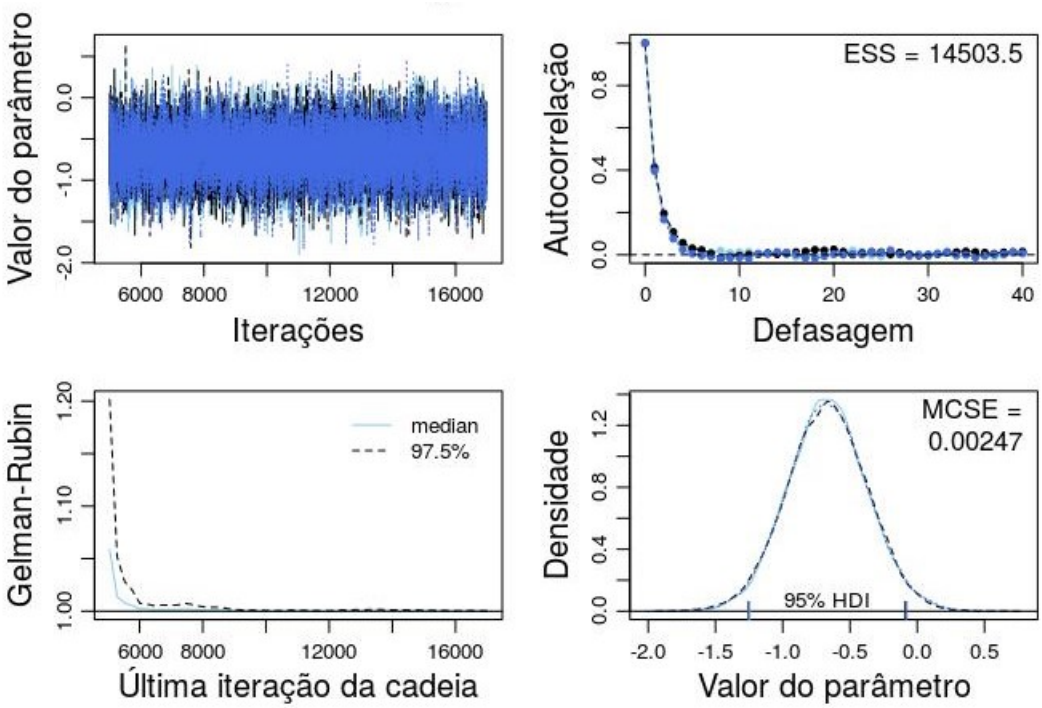


Figura 34 – Diagnóstico MCMC para o parâmetro δ do modelo COM-Poisson AR(1) ZM para os dados de SRC

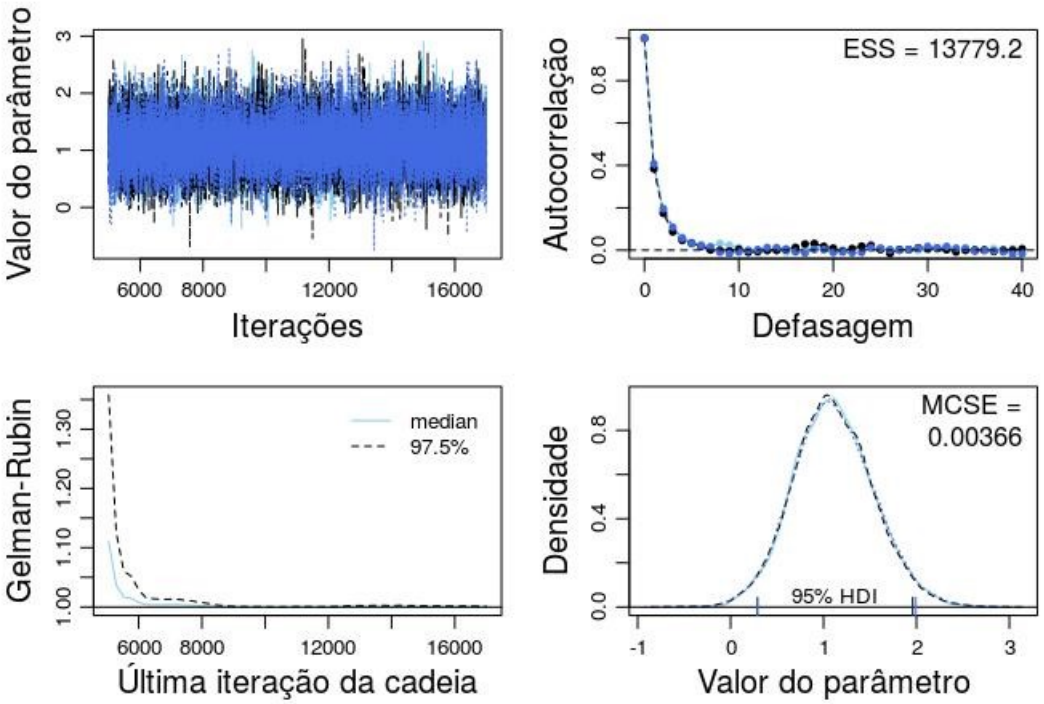
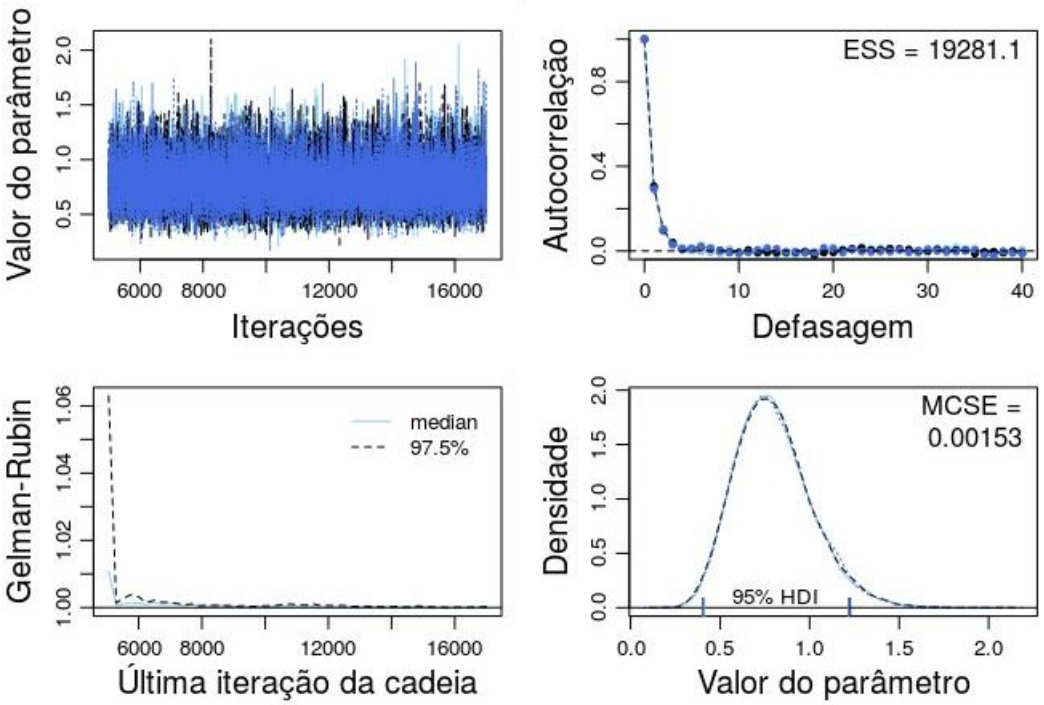
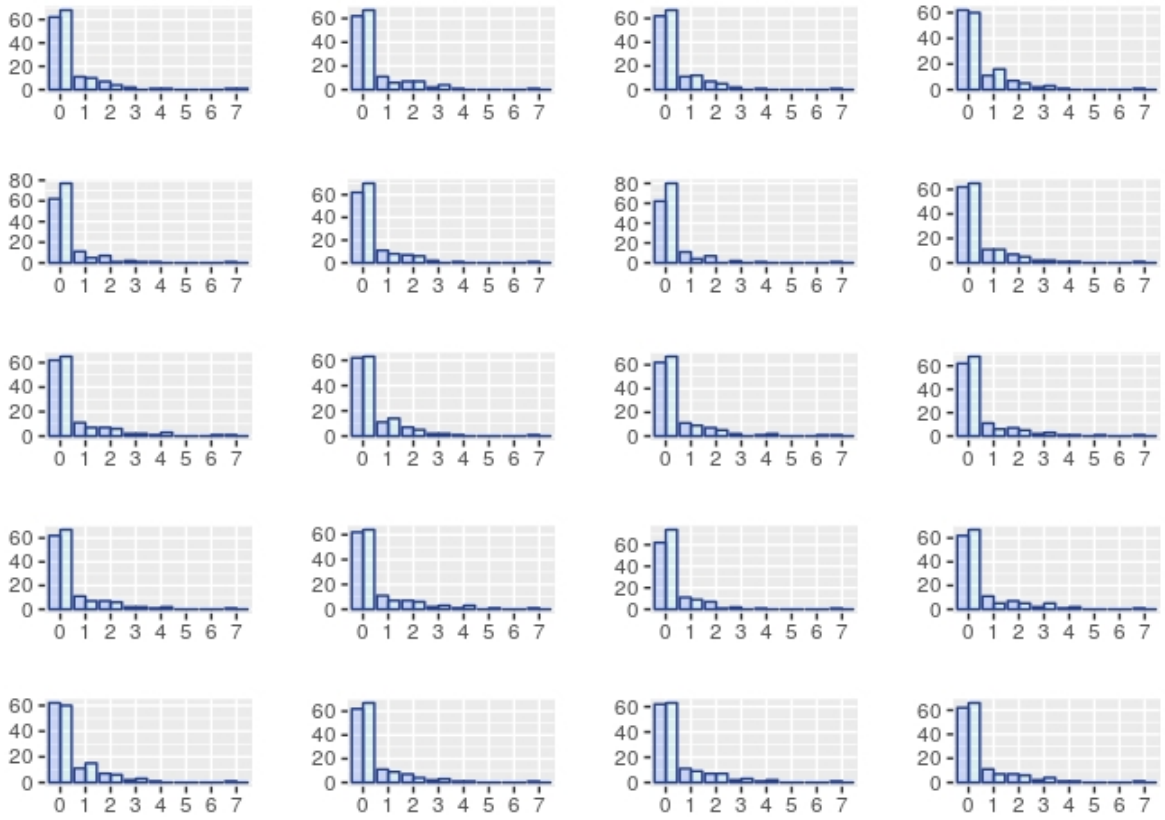


Figura 35 – Diagnóstico MCMC para o parâmetro φ do modelo COM-Poisson AR(1) ZM para os dados de SRC



Note, por meio dos intervalos de credibilidade na Tabela 11 e Figura 31, que todos os parâmetros do modelo COM-Poisson AR(1) ZM sem o parâmetro β são significativos.

Figura 36 – Comparação entre valores observados e ajustados para o modelo COM-Poisson AR(1) ZM usando os dados de SRC. As barras à esquerda representam os dados observados e à direita os dados ajustados



A Figura 36 mostra o resultado do procedimento de checagem preditiva a posteriori para o modelo COM-Poisson AR(1) ZM. Foram gerados 20 conjuntos de dados fictícios do modelo COM-Poisson AR(1) ZM com 84 observações em cada conjunto.

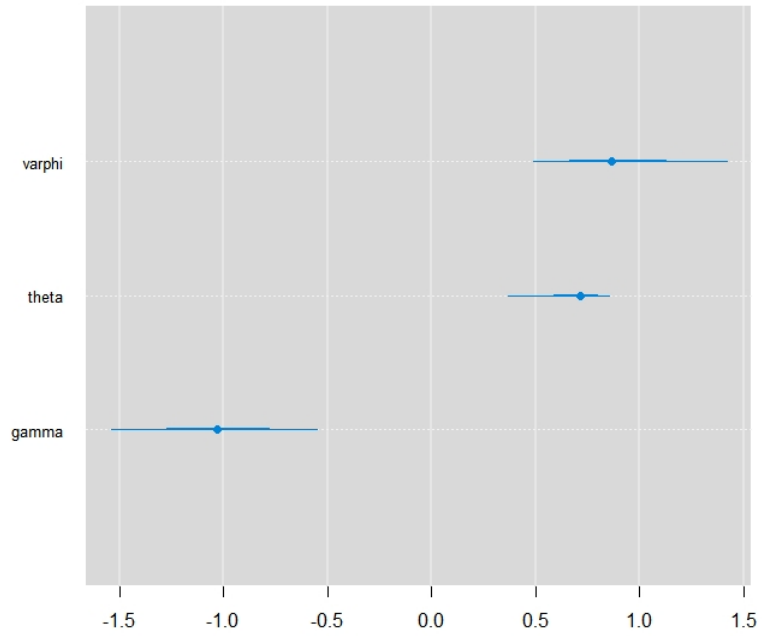
Para gerar as réplicas, foram amostradas 20 quádruplas dos parâmetros estimados pelo método MCMC, isto é, $\hat{\phi}$, $\hat{\gamma}$, $\hat{\delta}$ e $\hat{\varphi}$. Dos 20 conjuntos de dados fictícios do modelo COM-Poisson AR(1) ZM, 3 apresentaram obervações superiores a 5, o que não havia ocorrido com os modelos que utilizaram como base a distribuição de Poisson.

O modelo COM-Poisson MA(1) ZM foi ajustado aos dados de rubéola congênita e resultou em parâmetros significativos, como pode ser visto na Tabela 12 e na Figura 37.

Tabela 12 – Sumários das distribuições a posteriori para o modelo COM-Poisson MA(1) ZM para os dados de SRC

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
θ	$0,6877$	$0,1244$	$0,0012$	$[0,4293; 0,8784]$
γ	$-1,0333$	$0,2510$	$0,0017$	$[-1,5245; -0,5376]$
φ	$0,8873$	$0,2376$	$0,0018$	$[0,4499; 1,3592]$

Figura 37 – Intervalos de credibilidade para os parâmetros do modelo CP MA(1) ZM para os dados de SRC



Também foram estudadas outras ordens do modelo COM-Poisson ARMA ZM para os dados de rubéola congênita: AR(2) e ARMA(1,1), no entanto esses modelos resultaram em parâmetros não significativos.

5.3 Comparação entre os modelos analisados na aplicação com os dados de rubéola

Nesta seção, é feita uma comparação entre os modelos estudados para os dados de rubéola por meio do critério de informação da deviância.

A Tabela 13 mostra os valores de DIC para os modelos ajustados com as distribuições Poisson ARMA ZM e COM-Poisson ARMA ZM e cujos parâmetros foram todos significativos. Pode-se perceber que tanto para o modelo Poisson ARMA ZM como para o modelo COM-Poisson ARMA ZM, a melhor ordem foi AR(1).

Os valores de EBIC também foram calculados; no entanto, devido à similaridade com os valores de DIC, assim como não alteração nas conclusões, os valores de EBIC não foram replicados na Tabela 13.

Tabela 13 – Valores de DIC para os modelos Poisson ZM e COM-Poisson ZM ajustados aos dados de SRC

DIC	AR(1)	MA(1)
Poisson ZM	1674	1676
COM-Poisson ZM	1674	1677

5.4 Usando a distribuição Poisson ZM com os dados de sífilis

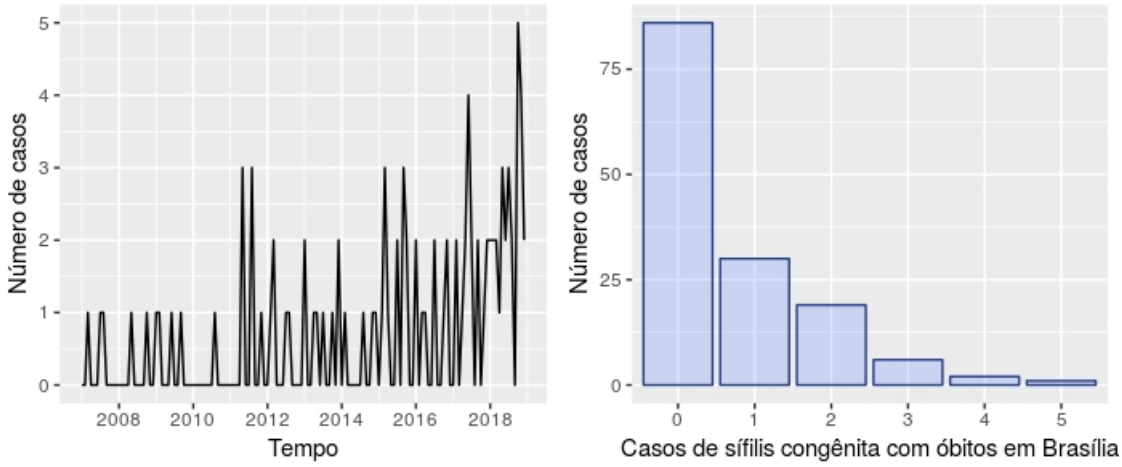
Para uma segunda aplicação, será utilizado o conjunto de dados de notificações de sífilis congênita em que houve complicações graves às crianças nascidas de mães com a doença. Foram selecionados os casos em que houve morte do bebê, seja por aborto espontâneo durante a gestação ou morte quando do nascimento, relacionados com a doença.

A sífilis, diferentemente da rubéola, não possui vacina disponível, sendo as melhores medidas para combatê-la a prevenção e acompanhamento durante o pré-natal caso a paciente esteja grávida. O tratamento é relativamente simples, mas se a doença não é tratada o feto pode sofrer complicações graves e chegar a óbito.

Apesar da quantidade de informação disponível em vários meios de comunicação, os casos de sífilis aumentaram no últimos anos e, sem um tratamento adequado ao tema por parte de gestores públicos, empresas e população, muitos casos que poderiam ser evitados ainda ocorrerão.

Os dados estudados são de Brasília-DF no período de janeiro de 2007 a dezembro de 2018, totalizando 144 observações. [2]

Figura 38 – Gráfico da série temporal e gráfico de barras referentes à sífilis congênita com óbito em Brasília-DF



[2] Os dados podem ser encontrados em <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/sifilisdf.def>, selecionando como linha o ano de diagnóstico e como coluna o respectivo mês, e a classificação final como "Natimorto/Aborto por Sífilis".

Para ilustrar a primeira aplicação com os dados de sífilis, foi escolhido primeiramente o modelo mais simples, isto é, Poisson autorregressivo de primeira ordem zero modificado. Os modelos utilizando os parâmetros β ou γ ou ambos resultaram em parâmetros não significativos, sendo escolhido então dentre os modelos Poisson AR(1) ZM aquele com apenas com os parâmetros ϕ e δ.

Os sumários a posteriori para os parâmetros são disponibilizados na Tabela 14 e os resultados de diagnóstico são apresentados nas Figuras 39 e 40.

Nesta aplicação foram utilizadas 3 cadeias, adaptação (*adapt*) feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1. Note que as cadeias convergiram e o tamanho amostral eficaz foi alto.

A Figura 41 mostra o resultado do procedimento de checagem preditiva a posteriori para o modelo Poisson AR(1) ZM. Foram gerados 20 conjuntos de dados fictícios do modelo Poisson AR(1) ZM com 144 observações em cada conjunto.

Para gerar as réplicas, foram amostrados 20 pares dos parâmetros estimados pelo método MCMC, isto é, $\hat{\phi}$ e $\hat{\delta}$. Com cada vetor de parâmetros estimado, foram geradas 20 amostras com 144 observações em cada. Nota-se que em alguns conjuntos de dados foi ajustada uma quantidade maior observações iguais a 1 (um) do que o observado nos dados reais.

Tabela 14 – Sumários das distribuições a posteriori para o modelo Poisson AR(1) ZM para os dados de sífilis

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
ϕ	$0,4504$	$0,1944$	$0,0013$	$[0,0712; 0,8269]$
δ	$0,5280$	$0,2560$	$0,0017$	$[0,0349; 1,0303]$

Figura 39 – Diagnóstico MCMC para o parâmetro ϕ do modelo Poisson AR(1) ZM para os dados de sífilis

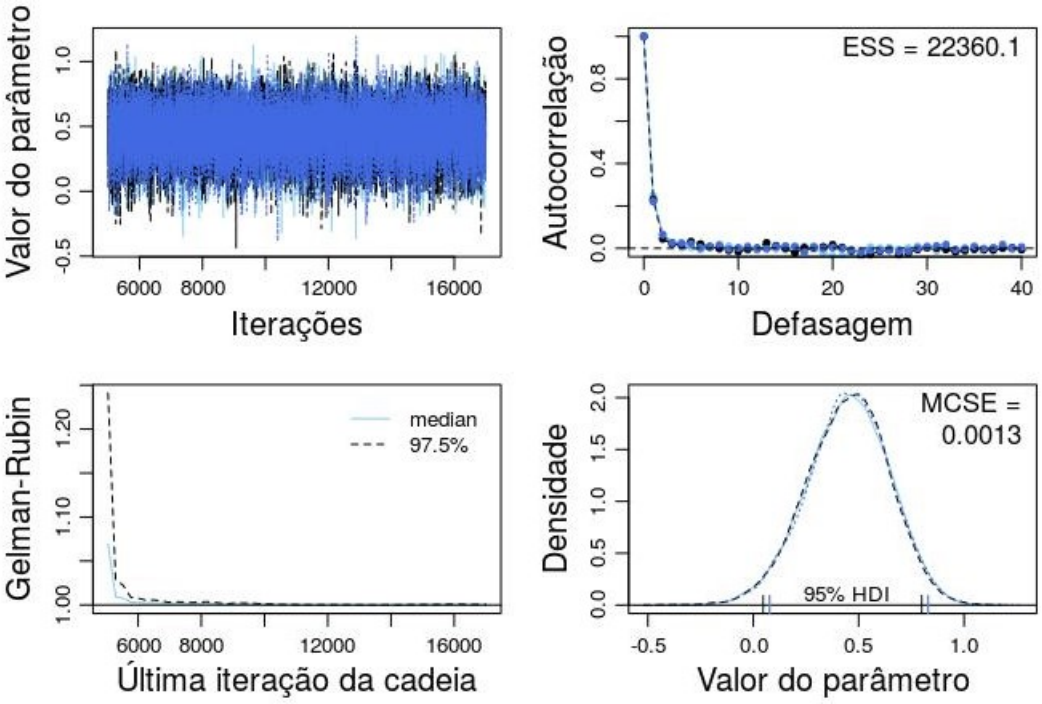
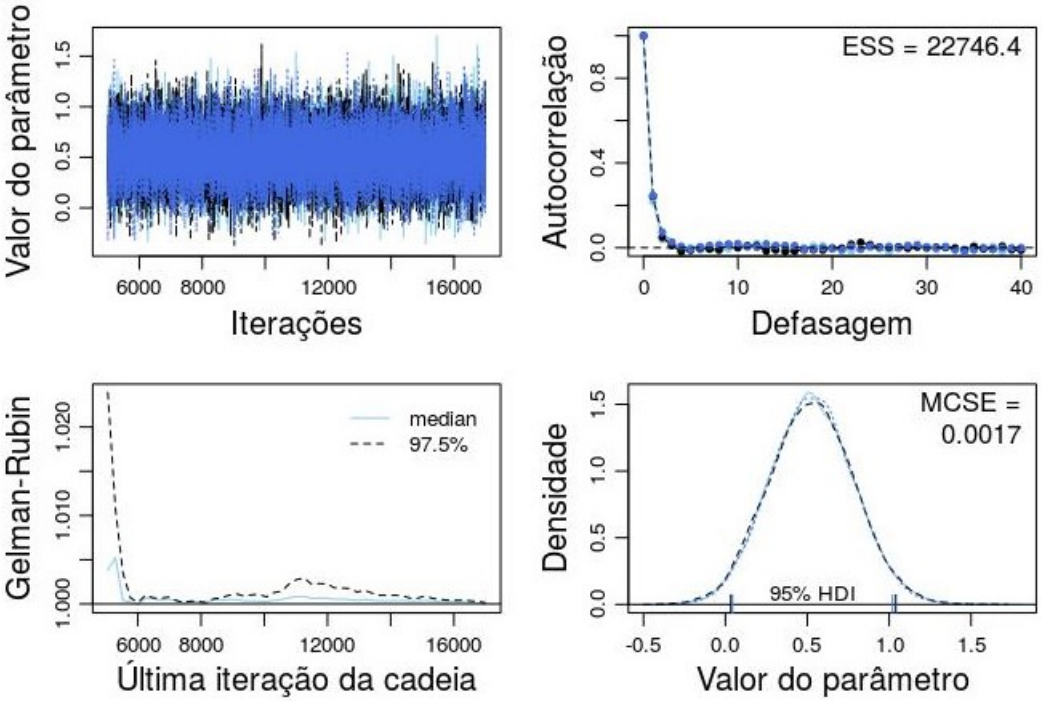


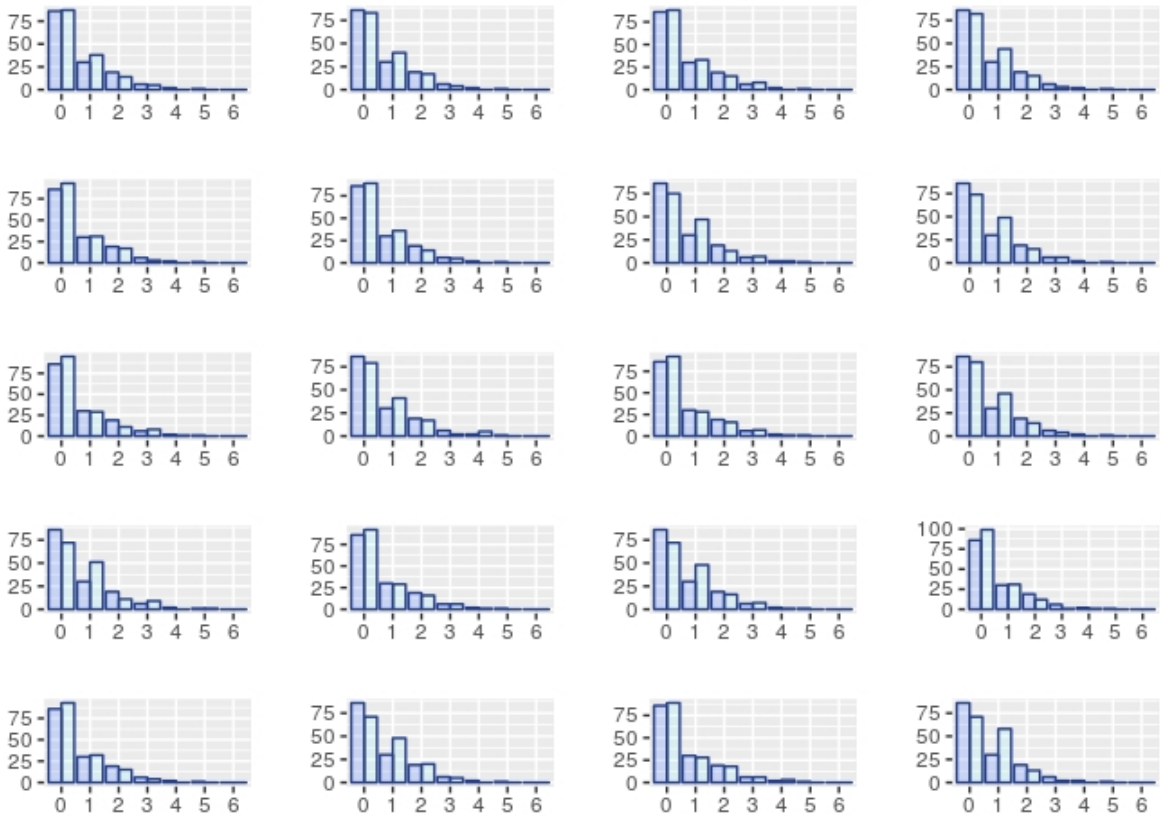
Figura 40 – Diagnóstico MCMC para o parâmetro δ do modelo Poisson AR(1) ZM para os dados de sífilis



O modelo Poisson AR(2) ZM também foi ajustado aos dados. No entanto, todos os parâmetros foram não significativos, como mostra a Figura 42, em que todos os parâmetros apresentaram regiões de equivalência prática incluindo o zero.

O modelo Poisson MA(1) ZM com os parâmetros θ e γ foi ajustado aos dados, e ambos

Figura 41 – Comparação entre valores observados e ajustados para o modelo Poisson AR(1) ZM usando os dados de sífilis. As barras à esquerda representam os dados observados e à direita os dados ajustados



os parâmetros foram significativos. O modelo com o parâmetro β, além dos parâmetros θ e γ, resultou em β não significativo.

Nesta aplicação também foram utilizadas 3 cadeias, adaptação feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1.

Assim, fazendo uso da equação (2.5), o modelo Poisson MA(1) ZM é escrito como:

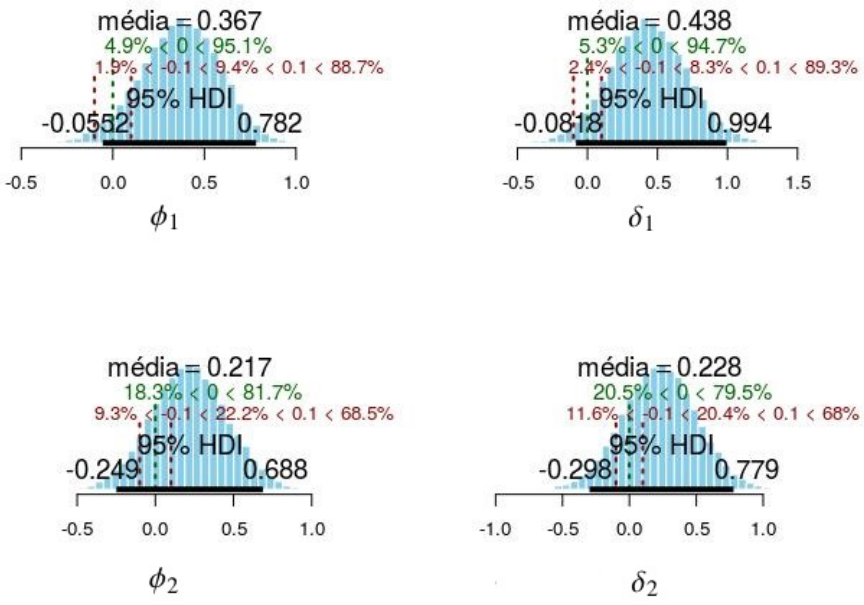
$$f_{ZMP}(y_t;\mu_t,\omega_t|\mathcal{F}_{t-1}) = (1-\omega_t)1_{(y_t)} + \omega_t f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}),$$

com $f_{ZTP}(y_t;\mu_t|\mathcal{F}_{t-1}) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!(1-e^{-\mu_t})}(1-1_{(y_t)})$. Para completar o modelo, as funções de ligação são dadas por:

$$\begin{aligned}\log(\mu_t) &= \theta\log\left(y^{\star}_{t-1}/\mu_{t-1}\right)\\ \text{logit}(\omega_t) &= \gamma.\end{aligned}$$

Os sumários a posteriori para os parâmetros são disponibilizados na Tabela 15 e os

Figura 42 – Distribuições a posteriori do modelo Poisson AR(2) ZM com regiões de equivalência prática para os dados de sífilis



resultados de diagnóstico são apresentados nas Figuras 43 e 44, que mostram que as cadeias convergiram e o tamanho amostral eficaz foi alto para ambos os parâmetros.

Nesta aplicação também foram utilizadas 3 cadeias, adaptação (*adapt*) feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1. Note que as cadeias convergiram e o tamanho amostral eficaz foi alto.

A medida DIC (critério de informação da deviância) foi calculada tanto para o modelo Poison AR(1) ZM quanto Poisson MA(1) ZM, e em ambos os modelos o valor foi igual a 2958.

Para verificar a capacidade do modelo Poisson MA(1) ZM de gerar dados próximos aos observados, a análise preditiva a posteriori foi realizada, usando a mesma metodologia dos modelos anteriores. O resultado encontra-se na Figura 45. As últimas barras de cada gráfico representam os valores iguais ou superiores a seis.

Tabela 15 – Sumários das distribuições a posteriori para o modelo Poisson MA(1) ZM para os dados de sífilis

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)
θ	$0,3402$	$0,1596$	$0,00108$	$[0,0236; 0,6311]$
γ	$-0,3854$	$0,1700$	$0,00114$	$[-0,7253; -0,0617]$

Figura 43 – Diagnóstico MCMC para o parâmetro θ do modelo Poisson MA(1) ZM para os dados de sífilis

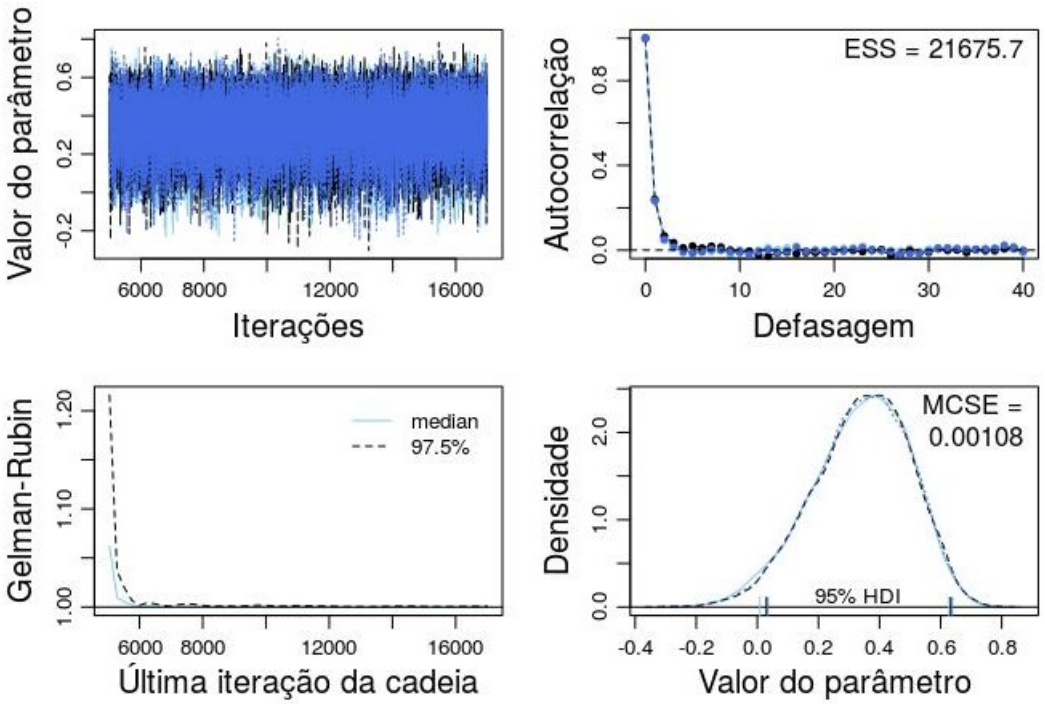
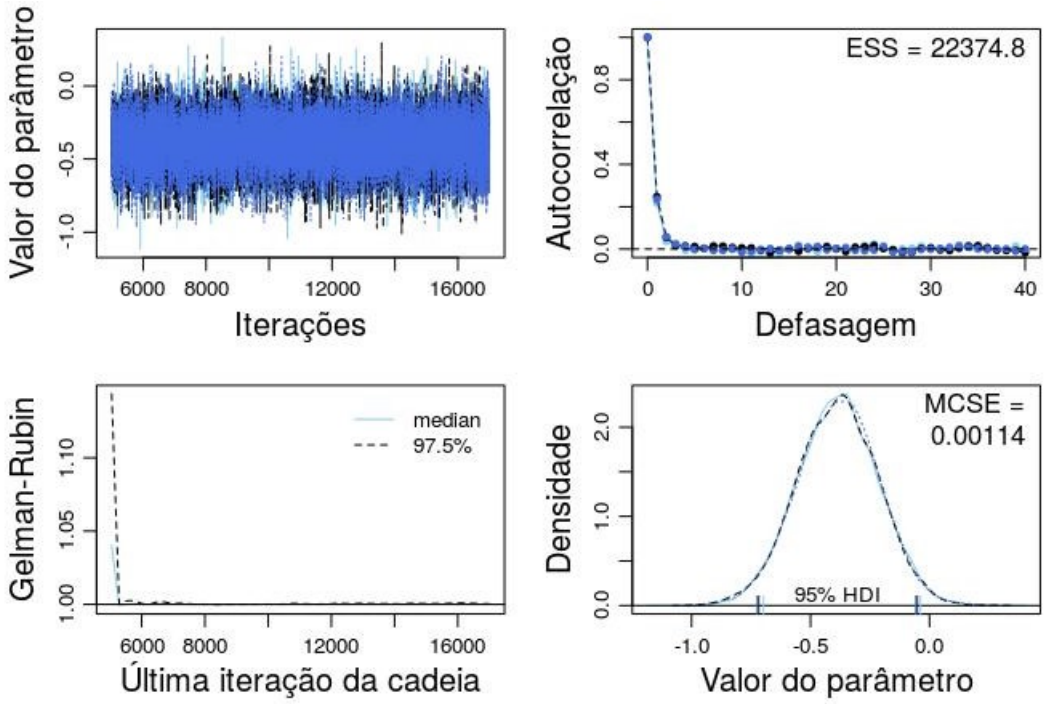
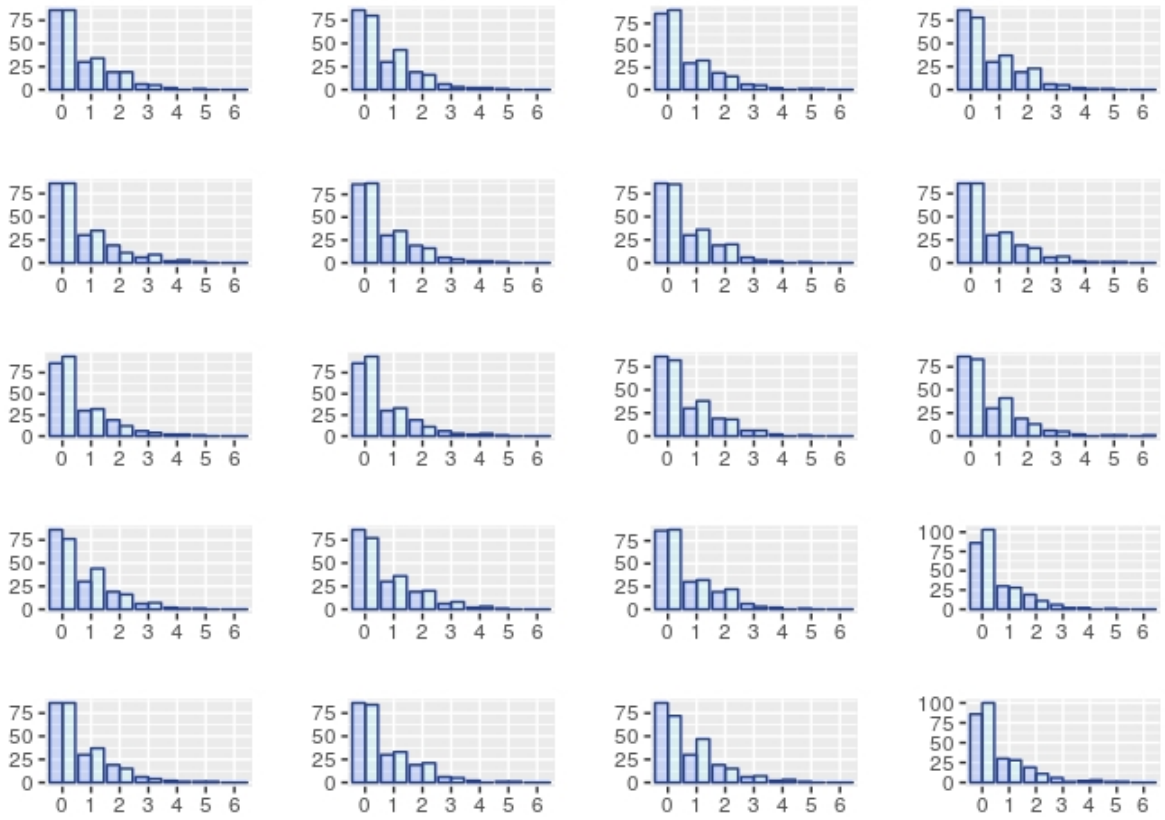


Figura 44 – Diagnóstico MCMC para o parâmetro γ do modelo Poisson MA(1) ZM para os dados de sífilis



O modelo Poisson ARMA(1,1) também foi ajustado aos dados de sífilis, no entanto, resultou em parâmetros não significativos.

Figura 45 – Comparação entre valores observados e ajustados para o modelo Poisson MA(1) ZM usando os dados de sífilis. As barras à esquerda representam os dados observados e à direita os dados ajustados



5.5 Usando a distribuição COM-Poisson ZM com os dados de sífilis

Fazendo uso agora da distribuição COM-Poisson, o modelo COM-Poisson AR(1) ZM com os parâmetros β e γ resultou em parâmetros não significativos, assim como o modelo com o parâmetro β e sem o parâmetro γ ou com o parâmetro γ, mas sem o parâmetro β.

O modelo sem os parâmetros β e γ resultou em todos os parâmetros serem significativos.

Nesta aplicação foram utilizadas 3 cadeias, adaptação feita em 1000 passos e *burning* de 4000 passos. O tamanho da amostra escolhido foi de 12000 para cada cadeia, com *thin* igual a 1.

Os sumários das distribuições a posteriori do modelo COM-Poisson AR(1) ZM para os dados de sífilis são dados na Tabela 16. As duas últimas colunas especificam o tamanho amostral eficaz e a estatística de Gelman-Rubin para cada parâmetro, respectivamente. O intervalo de credibilidade para o parâmetro φ inclui o valor 1, não excluindo, assim, a possibilidade de os dados terem como base a distribuição Poisson ZM.

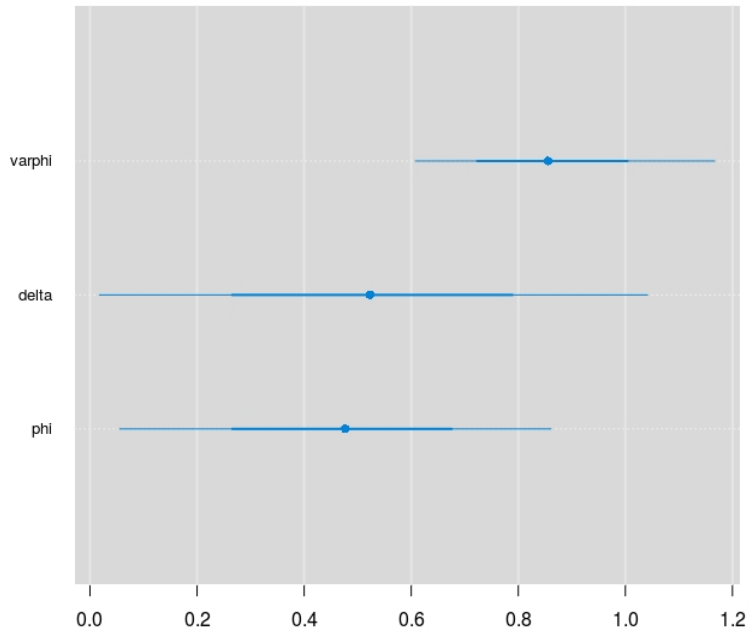
A Figura 46 ilustra os intervalos de credibilidade para os parâmetros do modelo. Os

intervalos interno e externo contêm 68% e 95% das observações, respectivamente. Vale ressaltar que esse gráfico mostra os intervalos de credibilidade utilizando os quantis da distribuição a posteriori de cada parâmetro, e não os intervalos de alta densidade.

Tabela 16 – Sumários das distribuições a posteriori para o modelo COM-Poisson AR(1) ZM para os dados de sífilis

Parâmetro	Média	Desvio Padrão	Erro Padrão	HDI (95%)	ESS	$\hat{R}$
ϕ	0,4713	0,2058	0,0014	[0,0689; 0,8732]	21701	1.0003
δ	0,5261	0,2629	0,0017	[0,0222; 1,0443]	22348	1.0001
φ	0,8636	0,1422	0,0010	[0,5935; 1,1498]	20259	1

Figura 46 – Intervalos de credibilidade para os parâmetros do modelo COM-Poisson AR(1) ZM para os dados de sífilis

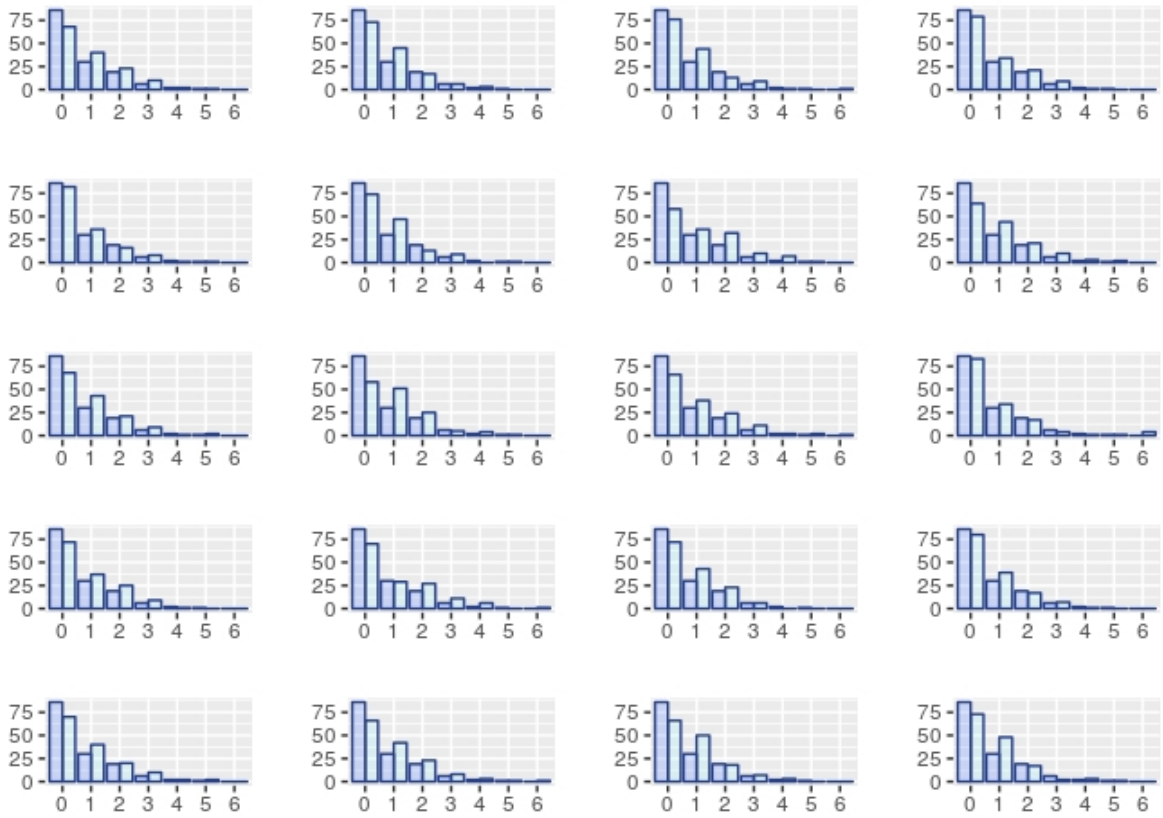


Por meio da Figura 46, depreende-se que todos os parâmetros são significativos, pois os intervalos de credibilidade não contêm o valor zero. Para verificar a capacidade do modelo COM-Poisson AR(1) ZM de gerar dados próximos aos observados, a análise preditiva a posteriori foi realizada, como mostra a Figura 47. As últimas barras de cada gráfico representam os valores a partir de seis.

O modelo COM-Poisson AR(2) também foi ajustado aos dados, porém resultou em parâmetros não significativos, mesmo retirando β e γ do modelo. A Figura 48 mostra esse resultado.

Os modelos COM-Poisson MA(1) ZM e ARMA(1,1) ZM também foram ajustados aos dados de sífilis, porém resultaram em parâmetros não significativos.

Figura 47 – Comparação entre valores observados e ajustados para o modelo COM-Poisson AR(1) ZM usando os dados de sífilis. As barras à esquerda representam os dados observados e à direita os dados ajustados



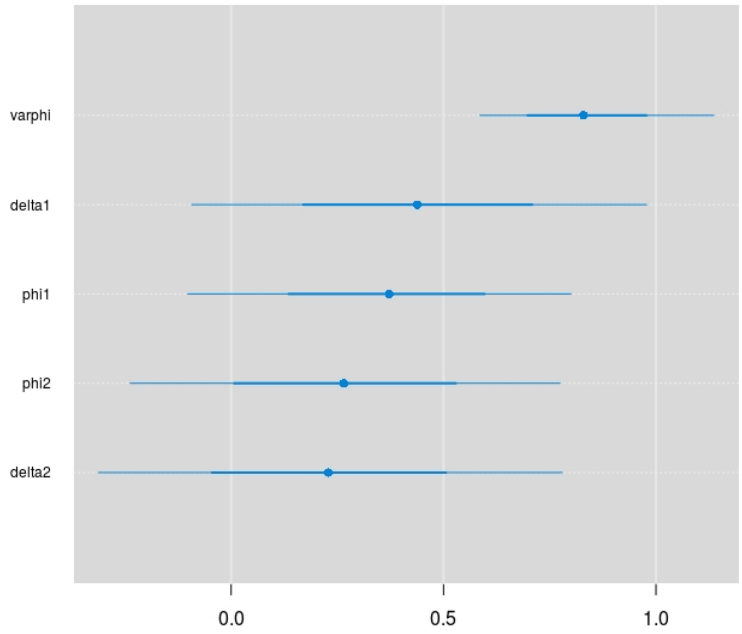
5.6 Comparação entre os modelos analisados na aplicação com os dados de sífilis

Nesta seção, é feita uma comparação entre os modelos estudados para os dados de sífilis por meio do critério de informação da deviância.

A Tabela 17 mostra os valores de DIC para os modelos ajustados com as distribuições Poisson ARMA ZM e COM-Poisson ARMA ZM. Pode-se perceber que para o modelo Poisson ARMA ZM, as melhores ordens foram AR(1) e MA(1), e para o modelo COM-Poisson ARMA ZM, a ordem escolhida foi AR(1), pois outras ordens estudadas levaram a modelos com parâmetros não significativos.

Os valores de EBIC também foram calculados; no entanto, devido à similaridade com os valores de DIC, assim como não alteração nas conclusões, os valores de EBIC não foram

Figura 48 – Intervalos de credibilidade para os parâmetros do modelo COM-Poisson AR(2) ZM para os dados de sífilis



replicados na Tabela 17.

Tabela 17 – Valores de DIC para os modelos Poisson ZM e COM-Poisson ZM ajustados aos dados de sífilis

DIC	AR(1)	MA(1)
Poisson ZM	2958	2958
COM-Poisson ZM	2958	-

CAPÍTULO

6

PREVISÃO EM MODELOS DA FAMÍLIA SÉRIE DE POTÊNCIA ARMA ZM

Neste capítulo são apresentados os resultados de previsão nos modelos Poisson AR(1) ZM e COM-Poisson AR(1) ZM para os dados de rubéola e sífilis estudados no capítulo 5.

6.1 Previsão no modelo Poisson ZM com os dados de SRC

A previsão para futuros valores do processo $Y(t+h)$ para algum $h > 0$, originando dos dados observados até o tempo t (D_t), é calculada pelo valor esperado condicional $\mathbb{E}(Y(t+h)|D_t)$ da densidade preditiva $f(y_{t+h}|D_t)$, dada por

$$f(y_{t+h}|D_t) = \int_{\boldsymbol{\Theta}_p} f(y_{t+h}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}|D_t) d\boldsymbol{\Theta}, \tag{6.1}$$

em que $\boldsymbol{\Theta}_p$ é o espaço paramétrico. Então a previsão $y(t+h)$, denotada por $\widehat{y}(t+h) = \mathbb{E}(Y(t+h)|D_t)$, é dada por

$$\widehat{y}(t+h) = \int_0^\infty y_{t+h} f(y_{t+h}|D_t) dy_{t+h}; \tag{6.2}$$

substituindo a equação (6.1) em (6.2) e alterando a ordem das integrais, encontra-se uma equação apropriada para calcular as predições, dada por

$$\widehat{y}(t+h) = \int_{\boldsymbol{\Theta}_p} \mathbb{E}(Y(t+h)|\boldsymbol{\Theta}, D_t) p(\boldsymbol{\Theta}|D_t) d\boldsymbol{\Theta}.$$

Estimativas de Monte Carlo para $\widehat{y}(t+h)$ podem ser calculadas considerando as amostras geradas pela distribuição a posteriori do vetor de parâmetros. Considerando o vetor de parâmetros providenciados pelo algoritmo MCMC, $\Theta^{(m)}, m = 1, \ldots, M$. Então, temos um vetor de valores $\mu^{(m)}(t+h), m = 1, \ldots, M$ calculados substituindo cada elemento do vetor $\Theta^{(m)}$ na função $\mu(t+h)$.

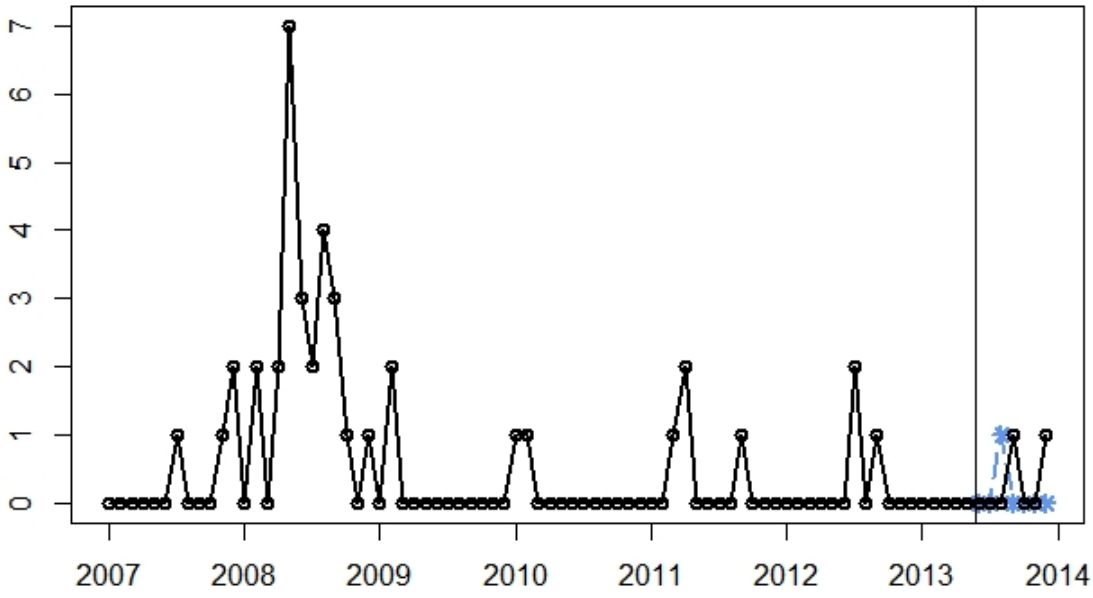
Uma estimativa de Monte Carlo para $\widehat{y}(t+h)$ pode ser calculada por:

$$\widehat{y}_{t+h} = \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}(t+h)$$

Para este estudo, foi considerado o modelo Poisson AR(1) ZM estudado na seção 5.1 do capítulo 5. Separando-se o conjunto de dados de $n = 84$ obervações em $\boldsymbol{y}'_{\text{treino}} = (y_1, \ldots, y_{78})$ e $\boldsymbol{y}'_{\text{teste}} = (y_{79}, \ldots, y_{84})$, o modelo Poisson AR(1) ZM foi implementado com os dados de treino e os dados de teste foram comparados com as previsões para as últimas seis observações.

Os resultados podem ser vistos na Figura 49. A linha vertical no gráfico marca o início

Figura 49 – Previsões para os dados de rubéola usando o modelo Poisson AR(1) ZM



da previsão. Na linha sólida estão os dados observados, e na linha tracejada estão as previsões. A Tabela 18 mostra os valores observados e os previstos para os dados de rubéola usando o modelo Poisson AR(1) ZM.

Tabela 18 – Previsões para os dados de rubéola usando o modelo Poisson AR(1) ZM

Horizonte	1	2	3	4	5	6
Valor Verdadeiro	0	0	1	0	0	1
Valor Previsto	0	1	0	0	0	0

6.2 Previsão no modelo COM-Poisson ZM com os dados de SRC

Para este estudo, foi considerado o modelo COM-Poisson AR(1) ZM estudado na seção 5.2 do capítulo 5, e a mesma separação dos dados da seção 6.1.

Os resultados podem ser vistos na Figura 50. A linha vertical no gráfico marca o início da previsão. Na linha sólida estão os dados observados, e na linha tracejada estão as previsões. A Tabela 19 mostra os valores observados e os previstos para os dados de rubéola usando o modelo COM-Poisson AR(1) ZM.

Figura 50 – Previsões para os dados de rubéola usando o modelo COM-Poisson AR(1) ZM

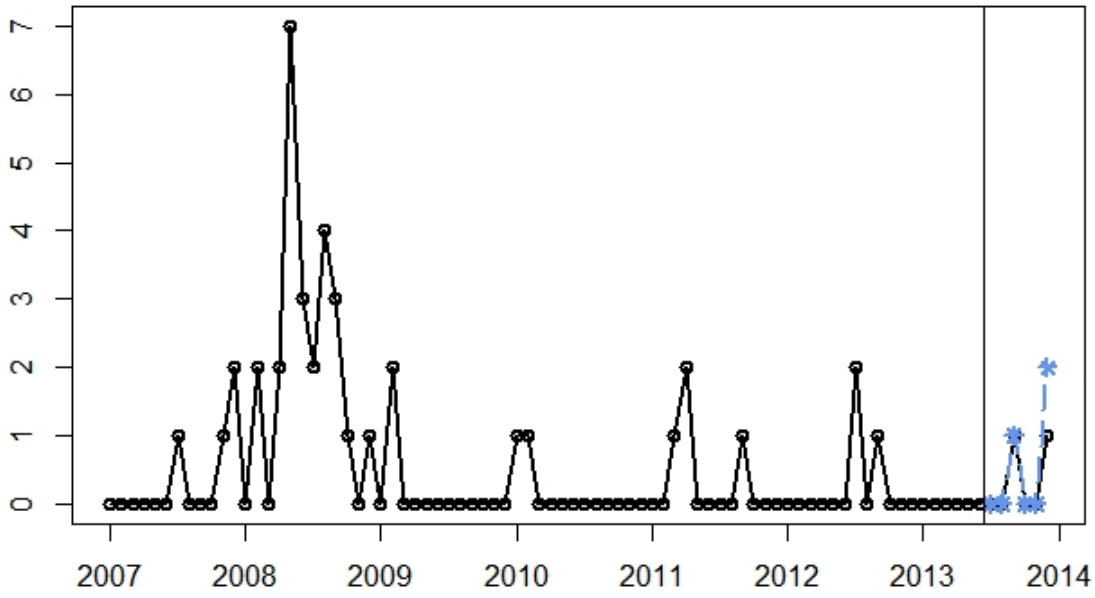


Tabela 19 – Previsões para os dados de rubéola usando o modelo COM-Poisson AR(1) ZM

Horizonte	1	2	3	4	5	6
Valor Verdadeiro	0	0	1	0	0	1
Valor Previsto	0	0	1	0	0	2

6.3 Erros de previsão dos modelos aplicados aos dados de SRC

Os erros de previsão são mostrados na Tabela 20. Nota-se que as medidas de acurácia RMSE e MAE indicam que o modelo COM-Poisson AR(1) ZM levou a erros de previsão levemente menores em comparação com o modelo Poisson AR(1) ZM.

Tabela 20 – Erros de previsão

Modelo	ME	RMSE	MAE
Poisson AR(1) ZM	0,1667	0,7071	0,5000
COM-Poisson AR(1) ZM	−0,1667	0,4082	0,1667

6.4 Previsão no modelo Poisson ZM com os dados de sífilis

Para este estudo, foi considerado o modelo Poisson AR(1) ZM estudado na seção 5.4 do capítulo 5. Separando-se o conjunto de dados de $n = 144$ obervações em $\boldsymbol{y}'_{\text{treino}} = (y_1, \ldots, y_{138})$ e $\boldsymbol{y}'_{\text{teste}} = (y_{139}, \ldots, y_{144})$, o modelo Poisson AR(1) ZM foi implementado com os dados de treino e os dados de teste foram comparados com as previsões para as últimas seis observações.

Os resultados podem ser vistos na Figura 51. A linha vertical no gráfico marca o início da previsão. Na linha sólida estão os dados observados, e na linha tracejada estão as previsões. A Tabela 21 mostra os valores observados e os previstos para os dados de sífilis usando o modelo Poisson AR(1) ZM.

Figura 51 – Previsões para os dados de sífilis usando o modelo Poisson AR(1) ZM

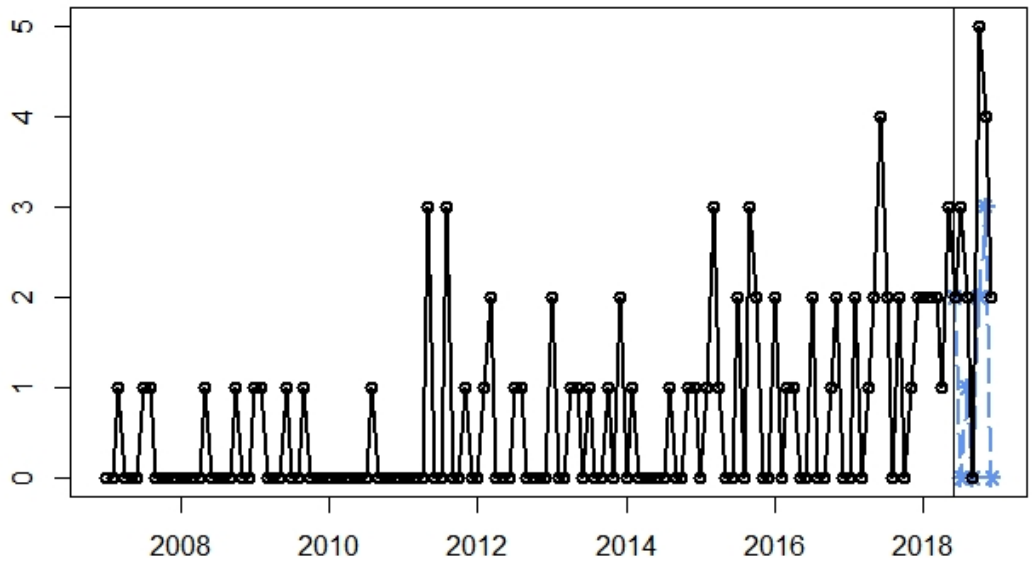


Tabela 21 – Previsões para os dados de sífilis usando o modelo Poisson AR(1) ZM

Horizonte	1	2	3	4	5	6
Valor Verdadeiro	3	2	0	5	4	2
Valor Previsto	0	1	0	2	3	0

6.5 Previsão no modelo COM-Poisson ZM com os dados de sífilis

Para este estudo, foi considerado o modelo COM-Poisson AR(1) ZM estudado na seção 5.5 do capítulo 5, e a mesma separação dos dados da seção 6.4.

Os resultados podem ser vistos na Figura 52. A linha vertical no gráfico marca o início da previsão. Na linha sólida estão os dados observados, e na linha tracejada estão as previsões. A Tabela 22 mostra os valores observados e os previstos para os dados de sífilis usando o modelo COM-Poisson AR(1) ZM.

Figura 52 – Previsões para os dados de sífilis usando o modelo COM-Poisson AR(1) ZM

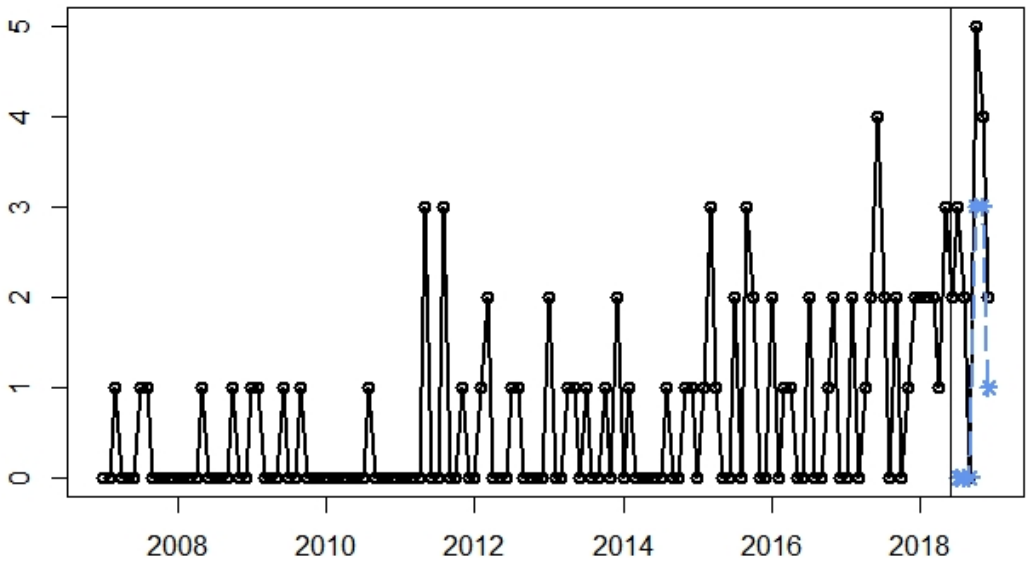


Tabela 22 – Previsões para os dados de sífilis usando o modelo COM-Poisson AR(1) ZM

Horizonte	1	2	3	4	5	6
Valor Verdadeiro	3	2	0	5	4	2
Valor Previsto	0	0	0	3	3	1

6.6 Erros de previsão dos modelos aplicados aos dados de sífilis

Os erros de previsão são mostrados na Tabela 23. Nota-se que as medidas de acurácia indicam que o modelo COM-Poisson AR(1) ZM levou a erros de previsão levemente menores em comparação com o modelo Poisson AR(1) ZM.

Tabela 23 – Erros de previsão

Modelo	ME	RMSE	MAE
Poisson AR(1) ZM	1,6667	2,0000	1,6667
COM-Poisson AR(1) ZM	1,5000	1,7795	1,5000

CAPÍTULO

7

CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Os modelos Poisson e COM-Poisson foram de rápida implementação - com modelos de ordens mais complexos levando um pouco mais tempo para serem implementados. Os modelos com base da distribuição Poisson ZM com ordens AR(1), AR(2), MA(1), ARMA(1,1) levaram 0.35, 2, 3 e 4 minutos, respectivamente. Os modelos com base da distribuição CP ZM com ordens AR(1), AR(2), MA(1), ARMA(1,1) levaram 2 minutos, 9 minutos, 8,7 horas e 8,5 horas, respectivamente.

Na aplicação em dados reais, foram analisados dois conjuntos de dados. Os modelos Poisson AR(1) ZM e COM-Poisson AR(1) ZM foram os mais competitivos, com a checagem preditiva a posteriori mostrando que os modelos conseguiram gerar dados com frequências próximas às dos observados e menores valores de DIC em comparação com outras ordens - MA(1) e ARMA(1,1).

Realizando um estudo de previsão com horizonte de seis observações usando os mesmos conjuntos de dados estudados na aplicação, foi constatado que ambos os modelos conseguiram prever com acurácia as observações, com erros levemente menores no modelo COM-Poisson AR(1) ZM em comparação com o modelo Poisson AR(1) ZM.

Assim, percebemos que os modelos estudados nesta tese são boas alternativas para ajustar dados com modificação no zero, com o modelo COM-Poisson ZM levando a previsões mais acuradas.

Como trabalhos futuros, são sugeridos estudos com séries temporais multivariadas. No contexto de dados de doenças, por exemplo, algumas enfermidades podem ocorrer concomitantemente, ou ainda podem estar relacionadas com o período de chuvas, por exemplo. Assim, o estudo de séries temporais com sazonalidade também ficou fora do escopo desta tese, podendo ser realizado em trabalhos futuros. Estudos utilizando outras funções de ligação também são

recomendados em novos trabalhos.

Os dados relacionados a doenças foram mais facilmente de serem obtidos para a realização desta tese, pois já há vários anos existe o DataSUS, repositório com dados de notificações de diversas condições clínicas. Esses dados possuem metodologia de registros que fazem com que seja possível buscar dados de qualquer região do país e realizar diversos filtros, possibilitando análises das mais diversas.

É possível que em alguns anos, com o conceito de dados abertos, outros conjuntos de dados sejam mais facilmente recuperados para fazer pesquisas. Casos de interesse público seriam os dados de segurança e dados jurídicos. Com a implementação do processo judicial eletrônico, espera-se que em alguns anos já haja um histórico suficiente para que sejam implementados estudos utilizando ferramentas de séries temporais para estudar e prever novos casos judiciais.

REFERÊNCIAS

ANDRADE, B. S.; ANDRADE, M. G.; EHLERS, R. S. Bayesian GARMA models for count data. *Communications in Statistics: Case studies, data analysis and applications*, v. 1, n. 4, p. 192–205, 2016. Citado na página 25.

BENJAMIN, M. A.; RIGBY, R. A.; STASINOPOULOS, D. M. Generalized autoregressive moving average models. *Journal of the American Statistical Association*, v. 98, p. 214–223, 2003. Citado nas páginas 25 e 41.

BERNARDO, J. M.; SMITH, A. F. M. *Bayesian Theory*. Chichester: John Wiley & Sons Ltd., 2000. (Wiley Series in Probability and Statistics). Disponível em: <https://cds.cern.ch/record/1319894>. Citado na página 45.

BOX, G. E. P.; TIAO, G. C. *Bayesian Inference in Statistical Analysis*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1973. Citado na página 45.

BROEK, J. van den. A score test for zero inflation in a Poisson distribution. *Biometrics*, v. 51, n. 2, p. 738–743, 1995. Citado na página 24.

BROOKS, S. P.; GELMAN, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, v. 7, n. 4, p. 434–455, 1998. Citado na página 44.

CARLIN, B. P.; LOUIS, T. A. *Bayesian methods for data analysis*. Boca Raton, FL: CRC Press, 2009. Citado nas páginas 45 e 46.

CONCEIÇÃO, K. S. *Modelos série de potência zero-modificados*. 120 f. Tese (Doutorado em Estatística) — Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, 2013. Citado nas páginas 28, 30, 32 e 39.

CONCEIÇÃO, K. S.; ANDRADE, M. G.; LOUZADA, F. On the zero-modified Poisson model: a Bayesian analysis and posterior divergence measure. *Computational Statistics*, v. 29, p. 959–980, 2014. Citado nas páginas 24 e 32.

CONSUL, P. C.; FAMOYE, F. *Lagrangian probability distributions*. Boston: Birkhäuser, 2006. Citado na página 38.

CONWAY, R.; MAXWELL, W. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, v. 12, n. 2, p. 132–136, 1962. Citado na página 34.

COX, D. R.; LEWIS, P. A. W. *The statistical analysis of series of events*. Londres: Methuen & Co LTD, 1966. Citado na página 30.

DAVIS, R. et al. *Handbook of discrete-valued time series*. Boca Raton: Taylor & Francis Group, 2015. Citado na página 25.

DENWOOD, M. J. runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models. *Journal of Statistical Software*, v. 71, n. 9, p. 1–25, 2016. Citado nas páginas 50, 53 e 60.

DIETZ, E.; BÖHNING, D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics & Data Analysis*, v. 34, p. 441–459, 2000. Citado na página 24.

FREEDMAN, L. S.; LOWE, D.; MACASKILL, P. Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, v. 40, p. 575–586, 1984. Citado na página 45.

GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, v. 85, n. 410, p. 398–409, 1990. Citado na página 43.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, v. 7, n. 4, p. 457–511, 1992. Citado na página 44.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analsysis and Machine Intelligence*, v. 6, p. 721–741, 1984. Citado na página 43.

GUIKEMA, S.; GOFFELT, J. A flexible count data regression model for risk analysis. *Risk Analysis*, v. 28, n. 1, p. 213–223, 2008. Citado nas páginas 34 e 35.

GUPTA, R. C. Modified power series distribution and some of its applications. *The Indian Journal of Statistics, Series B*, v. 36, n. 3, p. 288–298, 1974. Citado na página 28.

HALL, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, v. 56, p. 1030–1039, 2000. Citado na página 24.

HOBBS, B. P.; CARLIN, B. P. Practical bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, v. 18, p. 54–80, 2007. Citado na página 45.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, v. 22, p. 679–688, 2006. Citado na página 48.

KEDEM, B.; FOKIANOS, K. *Regression models for time series analysis*. Hoboken: John Wiley & Sons, 2002. Citado na página 41.

KRUSCHKE, J. K. Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, v. 6, n. 3, p. 299–312, 2011. Citado na página 45.

KRUSCHKE, J. K. *Doing Bayesian data analysis: A tutorial with R, JAGS and Stan*. Waltham, MA: Academic Press / Elsevier, 2015. Citado nas páginas 44, 45 e 62.

LAMBERT, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, v. 34, n. 1, p. 1–14, 1992. Citado na página 24.

LEE, A. H.; WANG, K.; YAU, K. K. W. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal*, v. 43, p. 963–975, 2001. Citado na página 24.

METROPOLIS, N. et al. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, v. 21, p. 1087–1091, 1953. Citado na página 44.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. Hoboken: John Wiley & Sons, 2006. Citado na página 30.

MÜLLER, J.; BOGENBERGER, K. Time series analysis of booking data of a free-floating carsharing system in berlim. *Transportation Research Procedia*, v. 10, p. 345–354, 2015. Citado na página 23.

PLUMMER, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: 3RD INTERNATIONAL WORKSHOP ON DISTRIBUTED STATISTICAL COMPUTING (DSC 2003), 2003, Viena, Austria. *Proceedings of the 3rd international workshop on distributed statistical computing*. [S.l.], 2003. p. 10. ISSN 1609-395X. Citado na página 43.

PLUMMER, M. *rjags: Bayesian Graphical Models using MCMC*. [S.l.], 2016. R package version 4-6. Disponível em: <https://CRAN.R-project.org/package=rjags>. Citado nas páginas 50, 53 e 60.

QUDDUS, M. A. Time series count data models: an empirical application to traffic accidents. *Accident Analysis and Prevention*, v. 40, p. 1732–1741, 2008. Citado na página 23.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado nas páginas 50, 53 e 60.

RODRIGUES, J. Bayesian analysis of zero-inflated distributions. *Communications in Statistics*, v. 32, n. 2, p. 281–289, 2003. Citado na página 24.

RUBIN, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, v. 12, n. 4, p. 1151–1172, 1984. Citado nas páginas 46 e 47.

SELLERS, K. F.; BORLE, S.; SHMUELI, G. The COM-Poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, v. 28, p. 104–116, 2012. Citado na página 34.

SHMUELI, G. et al. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, v. 54, p. 127–142, 2005. Citado na página 34.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc. B*, v. 64, n. 4, p. 583–639, 2002. Citado na página 46.

SPIEGELHALTER, D. J.; FREEDMAN, L. S.; PARMAR, M. K. B. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, p. 357–416, 1994. Citado na página 45.

SUNECHER, Y.; KHAN, N. M.; JOWAHEER, V. Estimation methods for a flexible INAR(1) COM-Poisson time series model. *Journal of Applied Mathematics, Statistics and Informatics*, v. 14, n. 1, p. 57–82, 2018. Citado na página 25.

YANG, M. *Statistical models for count time series with excess zeros*. 74 f. Tese (Doutorado em Bioestatística) — College of Public Health, University of Iowa, Iowa City, 2012. Citado nas páginas 23 e 25.

YAU, K. K. W.; LEE, A. H.; CARRIVICK, P. J. W. Modeling zero-inflated count series with application to occupational health. *Computer Methods and Programs in Biomedicine*, v. 74, p. 47–52, 2004. Citado na página 25.

ZEGER, S. L. A regression model for time series of counts. *Biometrika*, v. 75, n. 4, p. 621–629, 1988. Citado na página 23.

ZEGER, S. L.; QAQISH, B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, v. 44, p. 1019–1031, 1988. Citado na página 30.

APÊNDICE

A

CÓDIGOS UTILIZADOS NO CAPÍTULO 4

```
#### Poisson ARMA zero modificada ###
##################################################
########## Geração de dados artificias  ##########
##################################################

#Gerando n=156 observações de um modelo Poisson AR(1) ZM
n=156
rzmp=function(n,beta=0,phi,theta=0,gamma,delta,mu0=1,ys){
zmp<-c()
a=function(y){1/factorial(y)}
g=function(mu){mu}
f=function(mu) {exp(mu)}
r=length(phi);q=length(theta)
m=max(r,q)
N=n+100-m #descartar 100 primeiros devido a condicao inicial
for(i in m+1:N){
ystar=replace(x <- ys, x==0, 0.5)
mu=exp( beta + sum( phi*log( rev(ystar)[1:r] ) ) +
sum( theta * log( rev(ystar)[1:q] / rev(mu0)[1:q] ) )  ) #mu_t
k=exp( gamma + sum( delta * log( rev(ystar)[1:r] ) ) )
w=k/(1+k)                                    #w_t = omega_t
p=w/(1-exp(-mu))                             #p_t
y=0
p0=(1-p)+p*(1/f(mu))
u=runif(1)
Fy=p0
```

```
while(u>Fy){
y=y+1
Fy=Fy+p*(1/f(mu))*a(y)*(g(mu))^y
}
zmp[i]=y
ys=c(ys,y)
#   x0=c(1,runif(1))
mu0=c(mu0,mu)
}
return(zmp[-(1:100)])
}
set.seed(123)
amostra=rzmp(n,phi=0.8,gamma=-0.7, delta=1.3,ys=0)

table(amostra) #maior número amostrado foi "4"
mean(amostra) #0.3526
var(amostra) #0.5265

#GRÁFICOS
#install.packages("ggfortify")
#install.packages("forecast", dependencies=TRUE)
#install.packages("ggpubr")
#install.packages("reshape2")

library(ggfortify)
library(forecast)
library(ggpubr)
library(reshape2)

barplot(amostra)

##################################################
####### Análise exploratória dos dados ###########
##################################################

plot.ts(amostra)

library(astsa)
lag1.plot(amostra,12)
```

```
acf2(amostra)

library("ggfortify")
library(forecast)
require(gridExtra)

plot1 <- autoplot(as.ts(amostra),colour="black",ylab="",
xlab="Tempo", size=0.71)
plot2 <- ggplot(data=as.data.frame(amostra),  aes(amostra)) +
geom_bar(col="royalblue4", fill="royalblue1", alpha = .2) +
labs(x="Dados artificiais", y="Número de casos")
grid.arrange(plot1, plot2, ncol=2)


################################################
############### Análise Bayesiana ##############
################################################

#pacotes
#install.packages("R2jags", dependencies = TRUE,
repos = "http://cran.us.r-project.org")
#install.packages("runjags", dependencies = TRUE,
repos = "http://cran.us.r-project.org")
#install.packages("MCMCpack", dependencies = TRUE,
repos = "http://cran.us.r-project.org")
#install.packages("rjags")

library("R2jags")
library("runjags")
library("MCMCpack")
library("rjags")
library("coda")
#install.packages("DBDA2E-utilities.R")
#https://github.com/boboppie/kruschke-doing_bayesian_data_analysis/
blob/master/2e/DBDA2E-utilities.R
#baixe DBDA2E-utilities.R e salve na sua pasta de trabalho
# gráficos do livro "Doing Bayesian Data Analysis", de Kruschke
source("DBDA2E-utilities.R")
```

```
####### Ajustando modelo Poisson  AR(1) ZM #######
# Construindo o nome dos arquivos
fileNameRoot = paste0("PoissonAR1_phi_gamma_delta_artificiais")
saveType = "jpg"

y<-amostra
N<-156

dataList = list(
y = y,
N = N,
ones = rep(1,length(y)),
C = 10000
)

PoissonAR1_phi_gamma_delta_artificiais <- "
model {
mu[1]<- 0
y_s[1]<- 0.5*equals(y[1],0)+y[1]
for (i in 2:N) {
#y estrela
y_s[i]<-0.5*(y[i]==0)+y[i]

# verossimilhança de modelos Hurdle
l[i] <- (1-w[i])*(y[i]==0)+w[i]*fztps[i]

# distribuição de Poisson truncada nos zeros
fztps[i] <- exp(-mu[i]+y[i]*log(mu[i])-logfact(y[i])-
log(1-exp(-mu[i])))*(1-(y[i]==0))

# funções de ligação
mu[i] <- exp(phi*log(y_s[i-1]))
logit(w[i]) <- gamma+delta*log(y_s[i-1])

# truque de uns
f[i] <-  l[i]/C
ones[i] ~ dbern(f[i])
}
```

```
# priori
phi~dnorm(0,1.0E-5)
gamma~dnorm(0,1.0E-5)
delta~dnorm(0,1.0E-5)
}
"

writeLines(PoissonAR1_phi_gamma_delta_artificiais,
con="PoissonAR1_phi_gamma_delta_artificiais.txt" )

inits1 <- list(phi=0.5, gamma=-0.8, delta=1.1,
.RNG.name="base::Super-Duper", .RNG.seed=1)
inits2 <- list(phi=0.9, gamma=-0.9, delta=1.4,
.RNG.name="base::Wichmann-Hill", .RNG.seed=2)
inits3 <- list(phi=0.7, gamma=-0.5, delta=1.2,
.RNG.name="base::Wichmann-Hill", .RNG.seed=2)

startTime = proc.time()
set.seed(123)
PoissonAR1_phi_gamma_delta_artificiais<-
run.jags(model="PoissonAR1_phi_gamma_delta_artificiais" ,
monitor=c("phi","gamma","delta"),
data=dataList ,
inits=list(inits1,inits2,inits3) ,
n.chains=3,
adapt=1000,
burnin=4000,
sample=12000,
thin=1,
summarise=TRUE ,
plots=FALSE)
stopTime = proc.time()
duration = stopTime - startTime
show(duration)

# sumários do modelo
PoissonAR1_phi_gamma_delta_artificiais

# do pacote rjags
```

```
codaSamples = as.mcmc.list( PoissonAR1_phi_gamma_delta_artificiais )
save( codaSamples , file=paste0(fileNameRoot,"Mcmc.Rdata") )
mcmcMat = as.matrix( codaSamples )

# examinando as cadeias
# diagnósticos de convergência
diagMCMC( codaObject=codaSamples , parName="phi" ,
saveName=fileNameRoot , saveType=saveType )
diagMCMC( codaObject=codaSamples , parName="gamma" ,
saveName=fileNameRoot , saveType=saveType )
diagMCMC( codaObject=codaSamples , parName="delta" ,
saveName=fileNameRoot , saveType=saveType )

extract.runjags(PoissonAR1_phi_gamma_delta_artificiais, 'dic')
# DIC -> 3080
DIC=3080
penalty= 6.489e-05
EBIC=DIC+penalty*(log(N)-1) #EBIC=3080

library("mcmcplots")
par(mfrow=c(1,1))
caterplot(codaSamples)

#resultados numéricos
summary(codaSamples)
#Iterations = 5001:17000
#Thinning interval = 1
#Number of chains = 3
#Sample size per chain = 12000
#1. Empirical mean and standard deviation for each variable,
#plus standard error of the mean:
#  Mean     SD Naive SE Time-series SE
#phi    0.5691 0.2540 0.001339       0.001675
#gamma -0.5196 0.2556 0.001347       0.002530
#delta  1.5423 0.4090 0.002156       0.004041
#2. Quantiles for each variable:
#  2.5%     25%     50%     75%     97.5%
#phi    0.04836  0.4019  0.5784  0.7446  1.041742
#gamma -1.01457 -0.6887 -0.5237 -0.3516 -0.007755
```

```
#delta  0.76728  1.2636  1.5361  1.8123  2.374170

#data frame dos resultados a posteriori
out <- do.call(rbind.data.frame, codaSamples)
head(out)

# Intervalos HDI
HDIofMCMC = function( sampleVec , credMass=0.95 ) {
#Computes highest density interval from a sample of
#representative values, estimated as shortest credible interval.
#Arguments:#sampleVec is a vector of representative values
#from a probability distribution.
#credMass is a scalar between 0 and 1, indicating the mass within
#the credible interval that is to be estimated.
# Value:
#HDIlim is a vector containing the limits of the HDI
sortedPts = sort( sampleVec )
ciIdxInc = ceiling( credMass * length( sortedPts ) )
nCIs = length( sortedPts ) - ciIdxInc
ciWidth = rep( 0 , nCIs )
for ( i in 1:nCIs ) {
ciWidth[ i ] = sortedPts[ i + ciIdxInc ] - sortedPts[ i ]
}
HDImin = sortedPts[ which.min( ciWidth ) ]
HDImax = sortedPts[ which.min( ciWidth ) + ciIdxInc ]
HDIlim = c( HDImin , HDImax )
return( HDIlim )
}

# HDI's
posphi = out[,1]
posgamma = out[,2]
posdelta = out[,3]
round(HDIofMCMC(posphi,credMass=.95),4) # 0.0643 1.0542
round(HDIofMCMC(posgamma,credMass=.95),4) # -1.0283 -0.0240
round(HDIofMCMC(posdelta,credMass=.95),4) #  0.7506 2.3530

### Ajustando modelo Poisson  COM-Poisson AR(1)  ZM ###
# Construindo o nome dos arquivos
```

```
fileNameRoot = paste0("COMPoissonAR1_phi_gamma_delta_varphi_artificiais")
saveType = "jpg"

y<-amostra
N<-156

dataList = list(
y = y,
N = N,
ones = rep(1,length(y)),
C = 10000
)

COMPoissonAR1_phi_gamma_delta_varphi_artificiais <- "
model {
mu[1]<- 0
y_s[1]<- 0.5*equals(y[1],0)+y[1]
for (i in 2:N) {
#y estrela
y_s[i]<-0.5*(y[i]==0)+y[i]

# verossimilhança de modelos Hurdle
l[i] <- (1-w[i])*(y[i]==0)+w[i]*fztps[i]

# cálculo da função f(mu,varphi)
for (j in 1:100) {
S[i,j] <- exp(varphi*(j*log(mu[i]) - logfact(j)))
}

# distribuição de COM-Poisson zero-truncada
fztps[i] <- exp((varphi*y[i])*log(mu[i])-varphi*loggam(y[i]+1)-
log(1+sum(S[i,]) - 1))*(1-(y[i]==0))

# funções de ligação
mu[i] <- exp(phi*log(y_s[i-1]))
logit(w[i]) <- gamma+delta*log(y_s[i-1])

# truque de uns
f[i] <-  l[i]/C
```

```
ones[i] ~ dbern(f[i])
}

# priori
phi~dnorm(0,1.0E-5)
gamma~dnorm(0,1.0E-5)
delta~dnorm(0,1.0E-5)
varphi~dnorm(0,1.0E-5)
}
"

writeLines(COMPoissonAR1_phi_gamma_delta_varphi_artificiais,
con="COMPoissonAR1_phi_gamma_delta_varphi_artificiais.txt" )

inits1 <- list(phi=0.5, gamma=-0.8, delta=1.1, varphi=0.8,
.RNG.name="base::Super-Duper", .RNG.seed=1)
inits2 <- list(phi=0.9, gamma=-0.9, delta=1.4, varphi=0.9,
.RNG.name="base::Wichmann-Hill", .RNG.seed=2)
inits3 <- list(phi=0.7, gamma=-0.5, delta=1.2, varphi=1.2,
.RNG.name="base::Wichmann-Hill", .RNG.seed=2)

startTime = proc.time()
set.seed(123)
COMPoissonAR1_phi_gamma_delta_varphi_artificiais<-
run.jags(model="COMPoissonAR1_phi_gamma_delta_varphi_artificiais",
monitor=c("phi","gamma","delta","varphi"),
data=dataList ,
inits=list(inits1,inits2,inits3) ,
n.chains=3,
adapt=1000,
burnin=4000,
sample=12000,
thin=1,
summarise=TRUE ,
plots=FALSE)
stopTime = proc.time()
duration = stopTime - startTime
show(duration)
```

```
# sumários do modelo
COMPoissonAR1_phi_gamma_delta_varphi_artificiais

# do pacote rjags
codaSamples=
as.mcmc.list(COMPoissonAR1_phi_gamma_delta_varphi_artificiais)
save( codaSamples , file=paste0(fileNameRoot,"Mcmc.Rdata") )
mcmcMat = as.matrix( codaSamples )

# examinando as cadeias
# diagnósticos de convergência
diagMCMC( codaObject=codaSamples , parName="phi" ,
saveName=fileNameRoot , saveType=saveType )
diagMCMC( codaObject=codaSamples , parName="gamma" ,
saveName=fileNameRoot , saveType=saveType )
diagMCMC( codaObject=codaSamples , parName="delta" ,
saveName=fileNameRoot , saveType=saveType )
diagMCMC( codaObject=codaSamples , parName="varphi" ,
saveName=fileNameRoot , saveType=saveType )

extract.runjags(COMPoissonAR1_phi_gamma_delta_varphi_artificiais, 'dic')
# DIC -> 3079
DIC=3079
penalty=7.231e-05
EBIC=DIC+penalty*(log(N)-1) #EBIC=3079

library("mcmcplots")
caterplot(codaSamples)

# resultados numéricos
summary(codaSamples)
#Iterations = 5001:17000
#Thinning interval = 1
#Number of chains = 3
#Sample size per chain = 12000
#1. Empirical mean and standard deviation for each variable,
#plus standard error of the mean:
#           Mean     SD Naive SE Time-series SE
#phi     0.5460 0.2348 0.001238       0.001609
```

```
#gamma  -0.5249 0.2521 0.001329       0.002520
#delta   1.5376 0.4059 0.002139       0.004005
#varphi  1.3067 0.2682 0.001414       0.001853
#2. Quantiles for each variable:
#           2.5%     25%     50%     75%    97.5%
#phi     0.07057  0.3903  0.5536  0.7068  0.98842
#gamma  -1.01039 -0.6953 -0.5293 -0.3555 -0.02449
#delta   0.76704  1.2608  1.5308  1.8049  2.35681
#varphi  0.83833  1.1157  1.2890  1.4769  1.88347

#data frame dos resultados a posteriori
out <- do.call(rbind.data.frame, codaSamples)
head(out)

# HDI's
posphi = out[,1]
posgamma = out[,2]
posdelta = out[,3]
posvarphi = out[,4]
HDIofMCMC(posphi,credMass=.95) #0.08881518 1.00399774
HDIofMCMC(posgamma,credMass=.95) #-1.01820743 -0.03335083
HDIofMCMC(posdelta,credMass=.95) #0.7702424 2.3585947
HDIofMCMC(posvarphi,credMass=.95) #0.7906756 1.8207334
```


ICMC
USP
SÃO CARLOS