UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Natacha Neves

Modelos Lineares Generalizados para Dados de Contagem com Estrutura de Autocorrelação Temporal

Campina Grande - PB

Novembro de 2018

Natacha Neves

Modelos Lineares Generalizados para Dados de Contagem com Estrutura de Autocorrelação Temporal

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Ricardo Alves de Olinda

Campina Grande - PB

Novembro de 2018

N518m Neves, Natacha.

Modelos lineares generalizados para dados de contagem com estrutura de autocorrelação temporal [manuscrito] / Natacha Neves. - 2018.

43 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2018.

"Orientação : Prof. Dr. Ricardo Alves de Olinda , Coordenação do Curso de Estatística - CCT."

1. Modelos Lineares Generalizados. 2. Modelo GARMA. 3. Dados de Contagem. 4. Dados epidemiológicos. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

Natacha Neves

Modelos Lineares Generalizados para Dados de Contagem com Estrutura de Autocorrelação Temporal

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 29 de Novembro de 2018.

BANCA EXAMINADORA

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Prof. Dr. Kleber N. Nunes de Oliveira Barros
Universidade Estadual da Paraíba

Prof. Dr. Sílvio F. Alves Xavier Júnior
Universidade Estadual da Paraíba

Dedico este trabalho à minha querida Avó, Maria de Lourdes, que não se encontra mais conosco. Porém, me deixou palavras de incentivo, força e determinação. Enfim, agradeço imensamente a esta pessoa incrível que ela foi!

Agradecimentos

Agradeço a Deus por tudo que ele me tem feito até hoje. À minha família, especialmente à minha mãe e avó, por não terem desistido de mim e pelas palavras de conforto e força.

Ao professor e orientador, Dr. Ricardo Alves de Olinda, pela sua dedicação, paciência e, principalmente humildade. E, além disso, o agradeço por ter acreditado em mim.

Resumo

Modelos Autorregressivos e de Médias Móveis Generalizados (GARMA) são uma classe de modelos que foi desenvolvida para estender os modelos ARMA com distribuição Gaussiana para um contexto de séries temporais não Gaussianas. Neste trabalho estudou-se a classe GARMA para modelar séries temporais de dados de contagem com as distribuições de Poisson e Binomial Negativa. A principal finalidade foi apresentar novos procedimentos para a modelagem de séries temporais, que relacionassem os Modelos Lineares Generalizados (MLG) com estrutura Autorregressiva de Média Móvel (ARMA), a fim de modelar dados epidemiológicos de séries temporais (contagem). Para atingir tal finalidade, inicialmente, foram ajustados os modelos Poisson e Binomial Negativo com independência temporal, aplicando os MLG's e, posteriormente, ajustou-se esses modelos, porém com dependência temporal. Em ambos os casos, para escolha do modelo que melhor se ajustou aos dados, foi considerado o AIC, como critério de seleção. Os modelos com dependência temporal se mostraram mais eficientes por apresentarem AIC's bem menores. Portanto, os modelos GARMA Poisson e GARMA Binomial Negativo se mostraram mais eficientes para modelagem de dados de contagem com estrutura com dependência temporal.

Palavras-chave: Saúde Pública. Epidemiologia. Dados de Contagem. Modelo GARMA.

Abstract

Generalized Autoregressive moving average (GARMA) models are a class of models that was developed for extending the univariate Gaussian ARMA time series model to a versatile observation-driven model for non-Gaussian time series data. In this work it was studied the class GARMA to model time series counting data with Poisson and Negative Binomial distributions. The main goal was to present new procedures for the modeling of time series, which related Generalized Linear Models (GLM) with Autoregressive Moving Average (ARMA), in order to model epidemiological data of time series (counting). To achieve this purpose, initially, it was fitted the Poisson and Negative Binomial models with time independence, applying the GLM's and, posteriorly, these models were fitted, but with time dependence. In both cases, to choose the model that best fit the data, AIC was considered as the selection criterion. The models with dependence time were shown to be more efficient for presenting much smaller AIC's. Therefore, the GARMA Poisson and GARMA Negative Binomial models were more efficients for modeling time dependence counting data.

Key-words: Public Health. Epidemiology. Count Data. GARMA Model.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Séries temporais para dados de contagem são registros de frequência absoluta da ocorrência de determinados eventos em sucessivos intervalos de tempo, e tem como característica importante a dependência temporal entre observações. Elas surgem nas mais variadas áreas do conhecimento, tais como saúde pública, economia, indústria e fenômenos meteorológicos (FERREIRA, 2015). No contexto de Epidemiologia pode-se, por exemplo, estar interessado em modelar o número de crianças que sofrem de infecção respiratória aguda ao longo do tempo (SANTOS et al., 2017).

Os modelos de regressão para dados de contagem são muito utilizados nas mais variadas áreas do conhecimento (PAULA, 2013). Estes modelos integram um quadro especial de metodologias devido ao fato de a variável resposta tomar apenas valores inteiros não negativos. As distribuições Poisson e binomial, pertencentes à família exponencial, são as mais conhecidas, e as mais utilizadas para modelar dados de contagem. No entanto, sempre que existe sobredispersão, torna-se necessário recorrer a outras distribuições como, por exemplo, à distribuição binomial negativa.

É importante observar que, em algumas metodologias de análise, são requeridos alguns pressupostos que nem sempre são atendidos. Portanto, o estatístico não pode se omitir sob consequências valores elevados dos erros e inferências inconsistentes (viesadas). Com a utilização dos MLG's, os problemas com escalas, sobredispersão e excesso de zeros na variável resposta serão minimizados (ZEILEIS; KLEIBER; JACKMAN, 2007). No entanto, quando se tem um processo de contagem e este é quantificado ao longo do tempo, é provável que exista uma autocorrelação residual.

Uma opção para melhorar o ajuste, visto que o MLG não é capaz de capturar a dependência ao longo do tempo, seria o uso de modelos de séries temporais para dados de contagem. De acordo com Andrade (2013), uma extensão dos MLG's, inicialmente introduzidos por Nelder e Wedderburn (1972), foi proposta por (BENJAMIN; RIGBY; STASINOPOULOS, 2003), podendo ser utilizada na modelagem de tais séries temporais, proporcionando uma gama de opções na escolha da distribuição ideal. Uma vez que temos em mãos uma série temporal de contagem, uma distribuição de probabilidade conjunta discreta é necessária para um bom ajuste do modelo aos dados.

Diante do exposto, este trabalho teve como principal objetivo apresentar novos procedimentos para a modelagem de séries temporais, que relacionassem os Modelos Lineares Generalizados (MLG's) com estrutura Autorregressiva de Médias Móveis (ARMA), a fim de modelar dados epidemiológicos de séries temporais (contagem). Foi descrita e caracterizada a estrutura da classe dos modelos GARMA, destacando seus métodos de

estimação e análises de diagnósticos inerentes à análise de regressão; identificou-se a natureza das séries temporais provenientes de estudos epidemiológicos mediante o emprego de técnicas estatísticas de análises descritivas e de análises inferenciais; mostrou-se que o ajuste de regressão via modelos lineares generalizados para dados de contagem, com independência temporal, nem sempre é eficiente comparando-se com a classe de modelos GARMA e, além disso, apresentou-se novos modelos para dados de contagem na área de Saúde Pública, divulgando métodos que já existiam.

2 Revisão de Literatura

Neste capítulo concentram-se os principais aspectos da modelagem estatística para dados de contagem com estrutura de autocorrelação temporal, bem como uma revisão dos principais artigos que contribuíram para o desenvolvimento dessas teorias com aplicações nas mais variadas áreas do conhecimento, especificamente Saúde Pública. Posteriormente faz-se uma revisão dos principais artigos, teóricos e aplicados, que contribuíram para o surgimento da classe dos modelos GARMA.

2.1 Processo Linear

De acordo com Box e Jenkins (1976), o processo linear é baseado na ideia de que uma série temporal y_t com sucessivos valores altamente dependentes pode ser gerada de uma série de choques independentes ϵ_t. Estes choques têm uma distribuição usualmente normal e são chamados de ruídos brancos. Segundo Morettin e Toloi (2006)

$$E[\epsilon_t] = 0$$

e

$$\gamma_k = E[\epsilon_t, \epsilon_{t+k}] = \begin{cases} \sigma^2 & se \quad k = 0 \\ 0 & se \quad k \neq 0 \end{cases}$$

De acordo com Ferreira (2015), o ruído branco ϵ_t é transformado no processo y_t pelo que é chamado filtro linear $\varphi(B)$. A operação simplesmente é uma soma ponderada das observações anteriores, de modo que

$$y_t = \mu + \epsilon_t + \varphi_1\epsilon_{t-1} + \varphi_2\epsilon_{t-2} + \ldots = \mu + \varphi(B)\epsilon_t, \tag{2.1}$$

em que μ é a média do processo autorregressivo e B é o operador de defasagem de k períodos de tempo anterior, dado por

$$B^k\epsilon_t = \epsilon_{t-k}. \tag{2.2}$$

O modelo definido em (2.1) também pode ser escrito como a soma ponderada dos valores passados de y_t adicionando-se ϵ_t, então

$$y_t = \mu + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + \mu_t.$$

2.2 Modelagem Box-Jenkins

Para escolha de um determinado modelo, seria interessante que se compreenda totalmente o funcionamento de um certo fenômeno. Desta forma, seria possível escrever uma

expressão matemática que conseguiria explicar concretamente tal fenômeno. Não obstante, para a construção de um determinado modelo utiliza-se uma base de conhecimento empírico, teórico e matemático para estimação de parâmetros a partir dos dados experimentais (BOX et al., 2008). De acordo com Morettin e Toloi (2006), a metodologia Box e Jenkins, baseia-se em um determinado ciclo interativo para estruturar um modelo de previsão, fundamentado nos próprios dados da série. Tal ciclo interativo, é divido em quatro etapas:

1. Uma classe útil de modelos é considerada (especificação);
2. Como tal classe de modelos é extensa são desenvolvidas subclasses deste modelo, com intuito de identificar um modelo, baseado na análise de autocorrelações e autocorrelações parciais (identificação);
3. Estima-se os parâmetros do modelo identificado (estimação);
4. Faz-se a análise de resíduos e verifica-se se o modelo ajustado está adequado para fazer as previsões (verificação ou diagnóstico).

Caso o modelo escolhido não seja adequado para representar os dados, o ciclo deve ser repetido novamente, voltando a fase de identificação. Se o propósito é previsão, então devemos analisar diversos modelos, afim de escolher o melhor modelo ajustado com o menor erro quadrático médio de previsão. Evidentemente os modelos são postulados, pois são simples e com poucos parâmetros, contudo as previsões são bem precisas.

2.3 Modelo Autorregressivo de Média Móvel

O modelo Autorregressivo de Média Móvel (ARMA) proposto por Box e Jenkins (1976) é usado quando há autocorrelação entre as observações e autocorrelação entre os resíduos. A forma geral de um modelo ARMA é dado por

$$\tilde{y}_t = \phi_1\tilde{y}_{t-1} + \ldots + \phi_p\tilde{y}_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - \ldots - \theta_q\epsilon_{t-p}, \qquad t = 1, \ldots, n, \tag{2.3}$$

ou

$$\phi(B)\tilde{y}_t = \theta(B)\epsilon_t, \tag{2.4}$$

em que $\tilde{y}_t = y_t - \mu$, $\phi(B) = (1 - \phi_1 B^1 - \ldots - \phi_p B^p)$ e $\theta(B) = (1 - \theta_1 B^1 - \ldots - \theta_q B^q)$. De (2.4) tem-se que

$$\tilde{y}_t = \phi(B)\epsilon_t = \frac{\theta(B)}{\phi(B)}\epsilon_t.$$

O modelo definido na Equação 2.3 é conhecido como Autorregressivo de Média Móvel de ordem p e q, ou seja, ARMA(p,q).

Os modelos AR, MA e ARMA são utilizados apenas para séries temporais estacionárias, que não possuem raiz unitária. Porém, caso uma série seja não estacionária (possui raiz unitária), é possível diferenciá-la d vezes, até torná-la estacionária, então aplicar o modelo Autorregressivo Integrado de Média Móvel, ou seja, ARIMA(p,d,q).

Box e Jenkins (1976) formalizam a teoria da utilização de componentes autorregressivos e de médias móveis na modelagem de séries temporais utilizando-se de duas ideias básicas na criação de sua metodologia de construção de modelos:

i) A parcimônia: utilização do menor número possível de parâmetros para se obter uma representação adequada do fenômeno em estudo;

ii) A interatividade: informação empírica analisada teoricamente e o resultado deste estágio é confrontado com a prática sucessivas vezes, até que o modelo obtido seja satisfatório.

A identificação do modelo ARMA(p,q) se faz por meio das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), que é a correlação entre duas observações seriais, eliminando a dependência dos termos intermediários (FERREIRA, 2015).

Após a identificação dos modelos, deve-se estimar os parâmetros. Neste caso, os parâmetros a serem estimados são $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)'$ e $\sigma^2 = Var(\epsilon_t)$. Os parâmetros do modelo podem ser estimados pelo método da máxima verossimilhança (MILHORANÇA, 2014).

Estes modelos são importantes para análise de séries temporais, e apesar de muitas extensões dos mesmos, eles ainda mostram-se úteis na modelagem de séries temporais estacionárias e não estacionárias (modelos ARIMA), principalmente quando considera-se que y_t pode assumir qualquer valor real. No entanto, neste trabalho o enfoque foi em séries temporais de contagem, ou seja, séries que somente assumem valores naturais, muito comum na área da Saúde Pública. Sendo assim, foram considerados modelos mais adequados para este tipo de dados, o modelo GARMA que é uma extensão do modelo ARMA.

2.4 Modelos Lineares Generalizados

A importância dos MLG's não é apenas de índole prática, do ponto de vista teórico a sua importância advém, essencialmente, do fato da metodologia destes modelos constituírem uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações. Nelder e Wedderburn (1972) mostraram que a maioria dos problemas estatísticos, que surgem nas áreas de agricultura, demografia, ecologia, economia, geografia, geologia, história, medicina, ciência política, psicologia, sociologia, industria, etc,

podem ser formulados, de uma maneira unificada, como modelos de regressão. Conforme Resende e Biele (2002), esses modelos envolvem uma variável resposta univariada, variáveis explicativas e uma amostra aleatória de n observações, sendo que:

i) A variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica (distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens).

ii) As variáveis explicativas entram na forma de um modelo linear (componente sistemático).

iii) A ligação entre os componentes aleatório e sistemático é feita por meio de uma função monótona e diferenciável.

2.5 Família exponencial uniparamétrica

De acordo com Cordeiro e Demétrio (2013), a família exponencial uniparamétrica é caracterizada por uma função (de probabilidade ou densidade) da forma

$$f(y;\theta) \quad = \quad h(y)exp[\eta(\theta)t(y) - b(\theta)], \tag{2.5}$$

em que as funções $\eta(\theta)$, $b(\theta)$, $t(y)$ e $h(y)$ assumem valores em subconjuntos dos reais. As funções $\eta(\theta)$, $b(\theta)$ e $t(y)$ não são únicas.

Várias distribuições importantes podem ser escritas na forma (2.5), tais como Poisson, binomial, normal, gamma, e gama inversa. Cordeiro (1995), apresenta 24 distribuições na forma (2.5). O suporte da família exponencial (2.5), isto é, $\{y; f(y;\theta) > 0\}$, não pode depender do parâmetro. Assim, a distribuição uniforme em $(0, \theta)$ não é um modelo da família exponencial. Pelo teorema da fatoração de Neyman-Fisher, a estatística $t(Y)$ é suficiente para θ (CORDEIRO; DEMÉTRIO, 2013).

A família exponencial na forma canônica é definida a partir de (2.5), considerando que as funções $\eta(\theta)$ e $t(y)$ são iguais à função identidade, de forma que

$$f(y;\theta) \quad = \quad h(y)exp[\theta y - b(\theta)]. \tag{2.6}$$

Na parametrização (2.6), θ é chamado de parâmetro canônico. O logaritmo da função de verossimilhança correspondente a uma única observação no modelo (2.6) é dado por

$$\ell(\theta) = \theta y - b(\theta) + log[h(y)]$$

e, portanto, a função escore $U = U(\theta) = d\ell(\theta)/d\theta$ resulta em $\mathbf{U} = y - b^{'}(\theta)$.

O fato simples de se calcularem momentos da família exponencial (2.6) em termos de derivadas da função $b(\theta)$ (denominada de função geradora de cumulantes) em relação ao parâmetro canônico θ é muito importante na teoria dos modelos lineares generalizados, principalmente, no contexto assintótico.

Suponha que $Y_1, \ldots, Y_n$ sejam n variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), seguindo (2.5). A distribuição conjunta de $Y_1, \ldots, Y_n$ é dada por

$$f(y_1, \ldots, y_n; \theta) \quad = \quad \left[\prod_{i=1}^{n} h(y_i)\right] exp\left[\eta(\theta)\sum_{i=1}^{n} t(y_i) - nb(\theta)\right]. \tag{2.7}$$

A Equação (2.7) implica que a distribuição conjunta $Y_1, \ldots, Y_n$ é, também, um modelo da família exponencial. A estatística suficiente é $\sum_{i=1}^{n} T(Y_i)$ e tem dimensão um qualquer que seja n.

Conforme Cordeiro e Demétrio (2013) é, geralmente, verdade que a estatística suficiente de um modelo da família exponencial segue, também, a família exponencial. Por exemplo, se $Y_1, \ldots, Y_n$ são variáveis aleatórias *i.i.d.* com distribuição de Poisson $P(\theta)$, então a estatística suficiente $\sum_{i=1}^{n} T(Y_i)$ tem, também, distribuição de Poisson $P(n\theta)$ e, assim, é um modelo exponencial uniparamétrico.

2.6 Componente aleatório

O componente aleatório de um modelo linear generalizado é definido a partir da família exponencial uniparamétrica na forma canônica (2.5) com a introdução de um parâmetro $\phi > 0$ de perturbação. Nelder e Wedderburn (1972) ao fazerem isso, conseguiram incorporar distribuições biparaméticas no componente aleatório do modelo. Sendo assim, tem-se que,

$$f(y; \theta, \phi) \quad = \quad exp\{\phi^{-1}[y\theta - b(\theta)] + c(y, \phi)\}, \tag{2.8}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Quando ϕ é conhecido, a família de distribuições (2.8) é idêntica à família exponencial na forma canônica (2.5). O valor esperado e a variância de Y com distribuição na família (2.8) são

$$E(Y) = \mu = b^{'}(\theta) \quad \text{e} \quad Var(Y) = \phi b^{''}(\theta)$$

Observa-se, então, que ϕ é um parâmetro de dispersão do modelo e seu inverso ϕ^{-1}, uma medida de precisão. A função que relaciona o parâmetro canônico θ com a média μ (inversa da função $b^{'}(\cdot)$) é denotada por $\theta = q(\mu)$. A função da média μ na variância é representada por $b^{''}(\theta) = V(\mu)$. Denomina-se $V(\mu)$ de função de variância. Observa-se que o parâmetro canônico pode ser obtido de $\theta = \int V^{-1}(\mu)d\mu$, pois $V(\mu) = d\mu/d\theta$.

É importante observar que se ϕ não for conhecido, a família (2.8) pode, ou não, pertencer à família exponencial biparamétrica. Para (2.8) pertencer à família exponencial biparamétrica quando ϕ é desconhecido, a função $c(y, \theta)$ deve ser decomposta, segundo Cordeiro e McCullagh (1991), como $c(y, \theta) = \phi^{-1}d(y) + d_1(y) + d_2(\phi)$. Esse é o caso das distribuições Poisson, binomial e binomial negativa.

Convém salientar que a especificação do componente aleatório requer a definição de uma distribuição de probabilidades apropriada à variável resposta. Essa definição deve ser baseada nas propriedades da variável aleatória em questão [1]. O conhecimento dos modelos probabilísticos disponíveis e de suas principais propriedades é fundamental para uma escolha adequada.

2.7 Modelo Autorregressivo e de Médias Móveis Generalizado (GARMA)

De acordo com Andrade (2013), seja $\{Y_t\}$ uma série temporal e $F_{t-1} = (x_1, ..., x_{t-1}, y_1, ..., y_{t-1}, \mu_1, ..., \mu_{t-1})$, sendo que Y_t tem distribuição na família exponencial dada por:

$$f(y_t|F_{t-1}) = \exp\left\{\frac{y_t\theta_t - b(\theta_t)}{\phi} + c(y_t, \phi)\right\}, \tag{2.9}$$

em que θ_t e ϕ são os parâmetros canônico e de dispersão, respectivamente. Além disso, $b(\cdot)$ e $c(\cdot)$ são funções específicas.

Como em modelos lineares generalizados (MLG) μ_t (média) está relacionada com o preditor linear η_t, então

$$\begin{aligned} g(\mu_t) &= \eta_t = \mathbf{X}'\beta + \tau_t \\ \tau_t &= \sum_{j=1}^{p} \phi\Omega(y_{t-j}, x_{t-j}, \beta) + \sum_{j=1}^{q} \theta\Phi(y_{t-j}, \mu_{t-j}), \end{aligned} \tag{2.10}$$

em que Ω e Φ são funções que representam os termos autorregressivos e de médias móveis.

Na prática, o preditor linear em (2.10) pode ser reescrito em uma forma mais simples dada por

$$\eta_t = x_t'\beta + \sum_{j=1}^{p} \phi_j\{g(y_{t-j}) - x_{t-j}'\beta\} + \sum_{j=1}^{q} \theta_j\{g(y_{t-j}) - \eta_{t-j}\}. \tag{2.11}$$

Os parâmetros p e q são identificados utilizando-se os critérios da escolha BIC ou AIC. Segundo Stasinopoulos et al. (2017), o modelo GARMA(p,q) é definido por uma

[1] (trata-se de uma variável aleatória discreta ou contínua? É razoável assumir que sua distribuição seja simétrica? Em qual conjunto numérico essa variável pode assumir valores?).

componente aleatória dada por (2.9) e uma componente sistemática ou preditor linear dada pela Equação (2.11).

Nesta pesquisa foram consideradas três importantes distribuições discretas para a classe de modelos GARMA: Poisson, binomial e binomial negativa. Foram apresentadas cada uma delas com sua respectiva densidade e preditores. As Equações (2.10) são importantes para a relação de interesse da distribuição da família exponencial e do preditor linear.

2.7.1 Modelo GARMA Poisson

Suponha que $y_t|F_{t-1}$ seja uma distribuição de Poisson com média μ_t, então

$$f(y_t|F_{t-1}) = exp\{y_t\ log(\mu_t) - \mu_t - log(y_t!)\}. \tag{2.12}$$

Aqui, $Y_t|F_{t-1}$ possui distribuição pertencente à família exponencial com $\phi = 1$, $\theta_t = log(\mu_t)$, $b(\theta_t) = exp(\theta_t)$, $c(y_t, \phi) = -log(y_t!)$ e $\nu(\mu_t) = \mu_t$.

A função de ligação canônica para este modelo é a função logarítmica (ANDRADE, 2017). Sendo assim, o preditor linear será dado por

$$log(\mu_t) = \beta_0 + \sum_{j=1}^{p} \phi_j\{log\ y^*_{t-j}\} + \sum_{j=1}^{q} \theta_j\{log(y^*_{t-1}) - log(\mu_{t-j})\}, \tag{2.13}$$

em que, $y^*_{t-j} = max(y_{t-j}, c)$, $0 < c < 1$.

2.7.2 Modelo GARMA Binomial

Suponha que $y_t|F_{t-1}$ seja uma distribuição binomial com média μ_t, então

$$f(y_t|F_{t-1}) = exp\left\{y_t\ log\left(\frac{\mu_t}{m-\mu_t}\right) + m\ log\left(\frac{m-\mu_t}{m}\right) + log\left(\frac{\Gamma(m+1)}{\Gamma(y_t+1)\Gamma(m-y_t+1)}\right)\right\}. \tag{2.14}$$

A função de ligação canônica para este modelo é a função logarítmica. Sendo assim, o preditor linear é dado por

$$log\left(\frac{\mu_t}{m-\mu_t}\right) = \beta_0 + \sum_{j=1}^{p} \phi_j\{log\ y^*_{t-j}\} + \sum_{j=1}^{q} \theta_j\{log(y^*_{t-j}) - log(\mu_{t-j})\}, \tag{2.15}$$

com $y^*_{t-j} = max(y_{t-j}, c)$, $0 < c < 1$ e m é conhecido.

2.7.3 Modelo GARMA Binomial Negativo

De acordo com Andrade (2017), se y_t é uma série temporal tal que $y_t|F_{t-1} \sim BN(k, \mu_t)$, então

$$f(y_t|F_{t-1}) = exp\left(y_t log\left\{\frac{\mu_t}{\mu_t+k}\right\} + klog\left\{\frac{k}{\mu_t+k}\right\} + log\left\{\frac{\Gamma(k+y_t)}{\Gamma(y_t+1)\Gamma(k)}\right\}\right), \tag{2.16}$$

pertence a família exponencial com k conhecido.

A função de ligação canônica para este modelo é a função logarítmica. Sendo assim, o preditor linear é dado por

$$log\left(\frac{\mu_t}{\mu_t + k}\right) = \beta_0 + \sum_{j=1}^{p} \phi_j\{logy_{t-j}\} + \sum_{j=1}^{q} \theta_j\{log(y_{t-j}) - log(\mu_{t-j})\}. \tag{2.17}$$

Os modelos que não levam em consideração a estrutura de da dependência temporal (com enfoque da classe GARMA) discutidos parecem ser mais flexíveis para modelar dados de contagem com uma estrutura de média móvel autorregressiva (BRIËT; AMERASINGHE; VOUNATSOU, 2013). Benjamin, Rigby e Stasinopoulos (2003) aplicaram um modelo estacionário GARMA a uma série temporal de casos de poliomielite com tendência sazonal, utilizando uma função seno-cosseno com uma combinação no ciclo anual e semi-anual. No entanto, se o componente sazonal é assumido como estocástico, o modelo GARMA apresentado por Benjamin, Rigby e Stasinopoulos (2003) não é apropriado. Além disso, muitas séries temporais de dados de contagem, incluindo os dados que foram utilizados nesta pesquisa, nem sempre são estacionários.

Por outro lado, Andrade, Andrade e Ehlers (2015) ao analisarem o número de internações por dengue e o número de óbitos no Brasil, concluíram que a classe de modelos GARMA proporcionou uma estrutura flexível para modelar séries temporais de contagem.

2.7.4 Componente Sazonal

Segundo Andrade (2017), os componentes sazonais do modelo podem ser representados como β_{S1} e β_{S2}, usando funções cosseno (cos) e seno (sen) respectivamente. Estes dois termos podem ser incluídos no preditor, assim

$$\begin{aligned} log(\eta_t) &= \beta_0 + \beta_{S1} cos\left(\frac{2\pi t}{12}\right) + \beta_{S2} sen\left(\frac{2\pi t}{12}\right) + \\ &+ \sum_{j=1}^{p} \phi_j\{logy_{t-j}\} + \sum_{j=1}^{q} \theta_j\{log(y_{t-j} - log(\mu_{t-j}))\}. \end{aligned} \tag{2.18}$$

2.7.5 Método de Estimação

O procedimento de ajuste e estimação dos modelos GARMA foi realizado pelo método de estimação de máxima verossimilhança (para mais detalhes ver Benjamin, Rigby e Stasinopoulos (2003)). O método de estimação é baseado nos procedimentos clássicos dos Modelos Lineares Generalizados.

Seja $\{y_t\}$ uma série temporal onde as Equações (2.9) e (2.11) são satisfeitas. O vetor de parâmetros é $\boldsymbol{\vartheta}' = (\boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\theta}')$, onde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)'$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. Para o procedimento de estimação, a função de máxima verossimilhança aproximada nas r primeiras observações é utilizada. $F_r = \{y_1, \ldots, y_r\}$, em que $r =$

$\max(p, q)$. A função de verossimilhança parcial pode ser calculada considerando que y_{t-1} e y_t são condicionalmente independentes. Assim

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}|F_n) &\propto \prod_{t=r+1}^{n} f(y_t|F_t) \\ &\propto \prod_{t=r+1}^{n} exp\left\{\frac{y_t g(\mu_t) - b(g^{-1}(\mu_t))}{\phi} + c(y_t, \phi)\right\} \end{aligned} \tag{2.19}$$

em que $g(\mu_t)$ é a função de ligação dada por

$$g(\mu_t) = x_t^{'}\beta + \sum_{j=1}^{p} \phi_j\{g(y_{t-1}) - x_{t-j}^{'}\} + \sum_{j=1}^{q} \theta_j\{g(y_{t-j}) - g(\mu_{t-j})\}. \tag{2.20}$$

De acordo com as Equações acima $t = r + 1, \ldots, n$. As Equações (2.19) e (2.20) não possuem uma solução analítica, havendo assim a necessidade de rotinas de otimizações numéricas para solução das equações.

A classe de modelos GARMA apresenta características assintóticas interessantes (BENJAMIN; RIGBY; STASINOPOULOS, 2003). Esta teoria funciona bem para grandes conjuntos de dados, no entanto, em dados reais discretos é comum lidar com pequenos conjuntos de dados. Sendo assim, métodos de reamostragem podem ser uma solução para problemas assintóticos (ANDRADE, 2017).

2.7.6 Predição dos modelos GARMA

O estimador $\hat{\eta}_t$, para $t = r + 1, \ldots, n$ é obtido por

$$\hat{\eta}_t = x_t\hat{\beta} + \sum_{j=1}^{p} \hat{\phi}_j\{g(y_{t-j}) - x_{t-j}\hat{\beta}\} + \sum_{j=1}^{q} \hat{\theta}_j\{g(y_{t-j}) - \hat{\eta}_{t-j}\}. \tag{2.21}$$

Usando a Equação (2.21), a média do processo é dada por $\hat{\mu}_t = g^{-1}(\hat{\eta}_t)$, para $t = r + 1, \ldots, n$.

O preditor de valores futuros y_{t+h} para $h > 0$ é dado por $\hat{y}_{n+h} = E[Y_{n+h}|F_{n+1}]$, onde a informação até n é conhecida. Segue que

$$F_{n+1} = \{x_{n+1}, x_n, \ldots, x_1, y_n, y_{n-1}, \ldots, y_1, \mu_n, \mu_{n-1}, \ldots, \mu_1\},$$

$\hat{y}_{t+h}$ é chamado de preditor com origem em n e horizonte h. A previsão para os modelos GARMA é realizada recursivamente pelo preditor linear de cada modelo (ANDRADE, 2017).

2.7.7 Teste de *Chow*

De acordo com Chow (1960), esse teste permite avaliar se os resultados dos conjuntos de dados, antes e depois das datas selecionadas, permanecem inalterados, ou seja, se não

apresentam mudanças estruturais. Exemplificando no caso da seleção de uma data específica, seja n o número de observações de uma amostra. Divide-se essa amostra em duas partes, com base na data de quebra selecionada. A primeira parte contém n_1 observações; a segunda, $n_2 = n - n_1$ observações. Em seguida, define–se como β_1, o parâmetro calculado para a primeira subamostra de n_1 observações e β_2, o parâmetro calculado para a segunda subamostra de n_2 observações. A hipótese nula de que os parâmetros são constantes ao longo do tempo será dada por:

$$\begin{cases} H_0 : \beta_1 = \beta_2 \\ H_1 : \beta_1 \neq \beta_2 \end{cases}$$

Se a hipótese nula (H_0) não for rejeitada, conclui-se que os parâmetros são constantes ao longo do tempo; se for rejeitada, conclui-se que os parâmetros são instáveis ao longo do tempo. Logo, a estatística F para testar a igualdade de médias será:

$$F = \frac{(S_0 - S_1 - S_2)/k}{(S_1 + S_2)/(n_1 + n_2 - 2k)} \sim F_{[k, n_1+n_2-2k]}$$

em que S_0 é a soma dos quadrados dos resíduos, considerando toda a amostra; S_1 e S_2 são, respectivamente, a soma dos quadrados dos resíduos das duas subamostras, com n_1 e n_2 observações; k é o número total de parâmetros.

2.7.8 Seleção de modelos

Em geral, o algoritmo de ajuste deve ser aplicado não a um modelo isolado, mas a vários modelos de um conjunto bem amplo que deve ser, realmente, relevante para a natureza das observações que se pretende analisar (DOBSON; BARNETT, 2011). Se o processo é aplicado a um único modelo, não levando em consideração possíveis modelos alternativos, existe o risco de não se obter um dos modelos mais adequados aos dados.

De acordo com Cordeiro e Demétrio (2013), os métodos sequenciais ("*stepwise methods*") foram substituídos por procedimentos ótimos de busca de modelos. O procedimento de busca examina, sistematicamente, somente os modelos mais promissores de determinada dimensão q e, baseado em algum critério, exibe os resultados de ajuste dos melhores modelos de q variáveis explanatórias, com q variando no processo de 1 até o tamanho p do subconjunto final de modelos considerados bons.

Na abordagem clássica dentre os critérios de seleção mais conhecidos e utilizados estão o AIC (*Akaike Criterion Information*) e BIC (*Bayesian Criterion Information*).

Akaike (1974) concluiu que o viés é dado assintoticamente por d, em que d é o número de parâmetros a serem estimados no modelo e definiu seu critério de informação como:

$$AIC = -2l(\hat{\theta}) + 2d.$$

Em que $l(\theta)$ denota o máximo da função de log-verossimilhança e d é o número de parâmetros a serem estimados. O critério BIC, proposto por Schwarz (1978), penaliza o modelo a ser escolhido mais fortemente do que o AIC e é definido por:

$$BIC = -2l(\hat{\theta}) + dlog(n),$$

como n sendo o número de observações da série.

2.7.9 Teste de Dickey-Fuller Aumentado

Considerando o modelo $Y_t = \rho Y_t + \epsilon_t$, em que ϵ_t é o ruído branco. A série Y_t é estacionária e descrita por um AR(1), se $|\rho| < 1$. Mas, se $\rho = 1$, a série Y_t é não estacionária. Dickey e Pantula (1987) propuseram o teste sobre a presença da raiz unitária na série Y_t, quando o processo gerador da série é expresso por um destes modelos:

$$\begin{aligned}
\Delta Y_t &= \alpha + \beta_t + \lambda_3 Y_{t-1} + \epsilon_t \\
\Delta Y_t &= \alpha + \lambda_2 Y_{t-1} + \epsilon_t \\
\Delta Y_t &= \lambda_1 Y_{t-1} + \epsilon_t
\end{aligned}$$

em que $\lambda_i = \rho - 1$, $\forall_i = 1, 2, 3$, α e β são constantes para serem estimadas.

Para o teste da raiz unitária de Dickey-Fuller, estima-se a seguinte autorregressão:

$$\begin{aligned}
\Delta Y_t &= (\rho - 1) Y_{t-1} + \epsilon_t \\
\Delta Y_t &= \gamma Y_{t-1} + \epsilon_t
\end{aligned}$$

na qual $\Delta = (Y_t - Y_{t-1})$ é o operador de diferença e $\gamma = \rho - 1$.

As hipóteses a serem testadas são apresentadas de acordo com a Tabela 1, onde $H_0 : \rho = 1$ equivale a $H_0 : \lambda = 0$.

Por fim, se ao menos uma das hipóteses apresentadas na Tabela 1, não for rejeitada, a série é não estacionária, possuindo pelo menos uma raiz unitária.

Neste caso, testaremos as seguintes hipóteses:

- $H_0 : \gamma = 0$, existe, pelo menos, uma raiz unitária e sua variável não é estacionária;
- $H_1 : \gamma < 0$, não existe raiz unitária e sua variável é fracamente estacionária.

Tabela 1 – Testes da Raiz Unitária de Dickey-Fuller.

Modelos	**Hipótese Nula**	**Regras de decisão**
1	$\lambda_3 = 0$	$\tau_3 >$ valor crítico $\Rightarrow H_0$ não é rejeitada
	$(\alpha, \beta, \lambda_3) = (0, 0, 0)$	$\delta_2 <$ valor crítico $\Rightarrow H_0$ não é rejeitada
	$(\alpha, \beta, \lambda_3) = (\alpha, 0, 0)$	$\delta_3 <$ valor crítico $\Rightarrow H_0$ não é rejeitada
2	$\lambda_2 = 0$	$\tau_2 >$ valor crítico $\Rightarrow H_0$ não é rejeitada
	$(\alpha, \lambda_2) = (0, 0)$	$\delta_1 <$ valor crítico $\Rightarrow H_0$ não é rejeitada
3	$\lambda_1 = 0$	$\tau_1 >$ valor crítico $\Rightarrow H_0$ não é rejeitada

Fonte: Adaptada de Dickey e Fuller (1979).

As estatísticas τ_1, τ_2, τ_3, δ_1, δ_2 e δ_3 que constam na Tabela 1 são obtidas das seguintes formas:

$$\tau_3 = \frac{\lambda_3}{\sigma_{\lambda_3}}$$

$$\delta_3 = \frac{SQR(1) - SQR(1)}{\frac{3SQR(1)}{n}}$$

$$\tau_2 = \frac{\lambda_2}{\sigma_{\lambda_2}}$$

$$\delta_2 = \frac{SQR(2) - SQR(2)}{\frac{3SQR(2)}{n}}$$

$$\tau_1 = \frac{\lambda_1}{\sigma_{\lambda_1}}$$

$$\delta_1 = \frac{SQR(3) - SQR(3)}{\frac{3SQR(3)}{n}}$$

Em que $SQR(1)$ é a soma de quadrado do resíduo do modelo $\Delta Z_t = \alpha + \beta_t + \lambda_3 Z_{t-1} + \epsilon_t$; $SQR(2)$ é a soma de quadrado dos resíduos do modelo$\Delta Z_t = \alpha + \lambda_2 Z_{t-1} + \epsilon_t$; $SQR(3)$ é a soma de quadrado dos resíduos do modelo $\Delta Z_t = \lambda_1 Z_{t-1} + \epsilon_t$, já σ_{λ_1} é a variância de λ_1; σ_{λ_2} é a variância de λ_2 e σ_{λ_3} é a variância de λ_3.

2.7.10 Teste de Mann-Kendall

De acordo com Hipel e McLeod (1994), o teste de Mann-Kendall é um teste estatístico não paramétrico para verificar se a tendência é estatisticamente significativa

ou não. Seja uma série temporal de observações $x_1, x_2, ..., x_n$, Mann (1945) propôs para a hipótese nula (H_0) que o dado vindo de uma população onde as variáveis aleatórias são independentes e igualmente distribuídas, entretanto, a hipótese alternativa (H_1) é que os dados seguem uma tendência monotônica no tempo. Sob H_0, o teste estatística de Mann-Kendall é:

$$S = \sum_{k=1}^{n_i-1} \sum_{j=k+1}^{n} sgn(x_j - x_k)$$

onde

$$sgn(x) = \begin{cases} +1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{2.22}$$

De acordo com a Equação (2.22), um valor positivo de S indica tendência positiva. Por outro lado, um valor negativo de S indica uma tendência negativa.

2.7.11 Análise de Intervenção

Em muitas aplicações ocorrem mudanças repentinas ou eventos extraordinários. Box e Tiao (1975) referem-se a tais eventos como intervenções.

Para Fokianos e Fried (2010) e Fokianos e Fried (2012) intervenções no modelo afetam a locação por incluir uma covariável determinística da forma $\delta^{t-\tau}\mathbb{1}(t \geq \tau)$, onde τ é o tempo de ocorrência e a taxa de decaimento δ é uma constante conhecida.

Esta definição pode ser aplicada a vários tipos de intervenções para diferentes escolhas da constante δ: um efeito singular para $\delta = 0$ (*spiky outlier*), um decaimento exponencial muda em locação para $\delta \in (0, 1)$ (mudança transitória) e uma mudança repentina de locação para $\delta = 1$ (mudança de nível). Similar para o caso das covariáveis, o efeito de uma intervenção é essencialmente aditiva para o modelo linear e multiplicativa para o modelo log-linear. Além disso, a intervenção entra na dinâmica do processo e, portanto, seu efeito no preditor linear não é puramente aditivo.

2.7.12 Regras de Pontuação (*Scoring Rules*)

De acordo com Gneiting, Balabdaoui e Raftery (2007), uma avaliação simultânea da calibração e nitidez resumidos em uma única pontuação (*score*) numérica pode ser realizada por regras de pontuação adequadas.

Uma pontuação é indicada pela distribuição preditiva P_t e a observação y_t por $s(P_t, y_t)$. A pontuação média para cada modelo correspondente é dado por $\sum_{t=1}^{n} \frac{s(P_t, y_t)}{n}$. Cada uma das diferentes regras de pontuações, nas quais, as mesmas sejam apropriadas

em sua aplicabilidade, capturam diferentes características da distribuição preditiva e a sua distância para os dados observados. Exceto pela pontuação dos erros normalizados, pois o modelo com menor pontuação é preferível.

O erro quadrático médio da pontuação é o único que não depende da distribuição e também é conhecido como erro quadrático médio de predição. A pontuação do erro quadrático médio normalizado mede a variância dos Resíduos de Pearson e é próximo de 1 se o modelo for adequado.

2.7.13 Transformação Integral da Probabilidade (*PIT*)

Uma ferramenta para avaliar a calibração probabilística da distribuição preditiva (ver Gneiting, Balabdaoui e Raftery (2007)) é a transformação integral da probabilidade, que seguirá uma distribuição uniforme se a distribuição preditiva estiver correta. Para dados de contagem Czado, Gneiting e Held (2009) define um valor PIT não aleatório para o valor observado y_t e a distribuição preditiva $P_t(y)$ por:

$$F_t(u|y) = \begin{cases} 0, & u \leq P_(y-1) \\ \frac{u-P_t(y-1)}{P_(y)-P_(y-1)}, & P_(y-1) < u < P_(y). \\ 1, & u \geq P_(y-1) \end{cases}$$

A média da transformação integral da probabilidade é dada por:

$$\overline{F}(u) = \frac{1}{n}\sum_{t=1}^{n} F_t(u|y_t), \quad 0 \leq u \leq 1.$$

2.7.14 *Bootstrap*

O método de simulação Bootstrap, proposto por DiCiccio e Efron (1996), se baseia na construção de distribuições amostrais por reamostragem, e é muito utilizado para estimar intervalos de confiança. Este método também pode ser utilizado, por exemplo, para estimar o viés e a variância de estimadores ou de testes de hipóteses calibrados. Além disso, o mesmo tem por base a ideia de que o pesquisador pode tratar sua amostra como se ela fosse a população que deu origem aos dados e usar a amostragem com reposição da amostra original para gerar pseudoamostras. A partir destas pseudoamostras é possível estimar características da população, tais como média, variância, percentis, etc.

Uma das formas de se obter amostras Bootstrap é o método não-paramétrico. Neste caso, cada amostra de tamanho n é obtida amostrando, com reposição, os dados originais, onde a estimação dos parâmetros é realizada para cada amostra, sendo este processo repetido B vezes. Na simulação não-paramétrica os dados não são gerados da distribuição de probabilidade dos dados, como no caso paramétrico.

Por exemplo, seja $x = (x_1, ..., x_n)$ uma amostra contendo n observações. Constrói-se então, B amostras $X^{*(1)}, ..., X^{*(B)}$ independentes, onde cada amostra é obtida a partir da reamostragem da amostra finita original finita inicial $x = (x_1, ..., x_n)$. Para cada uma das $X^{*(1)}, ..., X^{*(B)}$ amostras, estima-senos parâmetros de interesse (ANDRADE, 2013).

3 Resultados e Discussão

Neste capítulo segue uma aplicação com um conjunto de dados reais. O propósito geral deste capítulo é apresentar a viabilidade do Modelo GARMA, realizando um comparativo entre os ajustes dos modelos Poisson e Binomial Negativo, com independência e dependência temporal.

3.1 Número de casos notificados de crianças com Doenças Diarreicas Agudas

Os dados utilizados nesta pesquisa foram disponibilizados por Queiroz (2017), referente ao trabalho de dissertação intitulado "Doenças Diarreicas Agudas: ocorrência em crianças com idade entre zero a nove anos e a correlação com a qualidade da água". A pesquisa da referida autora foi realizada no município de Campina Grande entre os anos de 2006 a 2016, com dados do Açude Epitácio Pessoa (Boqueirão), Açude este que abastece a cidade de Campina Grande-PB. Por se tratar de dados de contagem, a referida autora utilizou como metodologia as distribuições Poisson e Binomial Negativa via Modelos Lineares Generalizados, ou seja, não foi considerada a estrutura de autocorrelação temporal. Sendo assim, como inovação dessa pesquisa, utilizou-se os modelos para dados de contagem com estrutura de autocorrelação temporal, ou seja, a classe dos modelos GARMA.

Inicialmente, tem-se o resumo das medidas descritivas da variável resposta e das variáveis explicativas. Por meio dessas medidas pode-se analisar o comportamento dos dados. Então, por meio da Tabela 1, observa-se que, em média, há aproximadamente 354 casos notificados de crianças com doenças diarreicas agudas (DDA) ao mês e, a princípio, existe uma grande variabilidade.

Tabela 2 – Estatísticas descritivas da variável resposta e das variáveis explicativas.

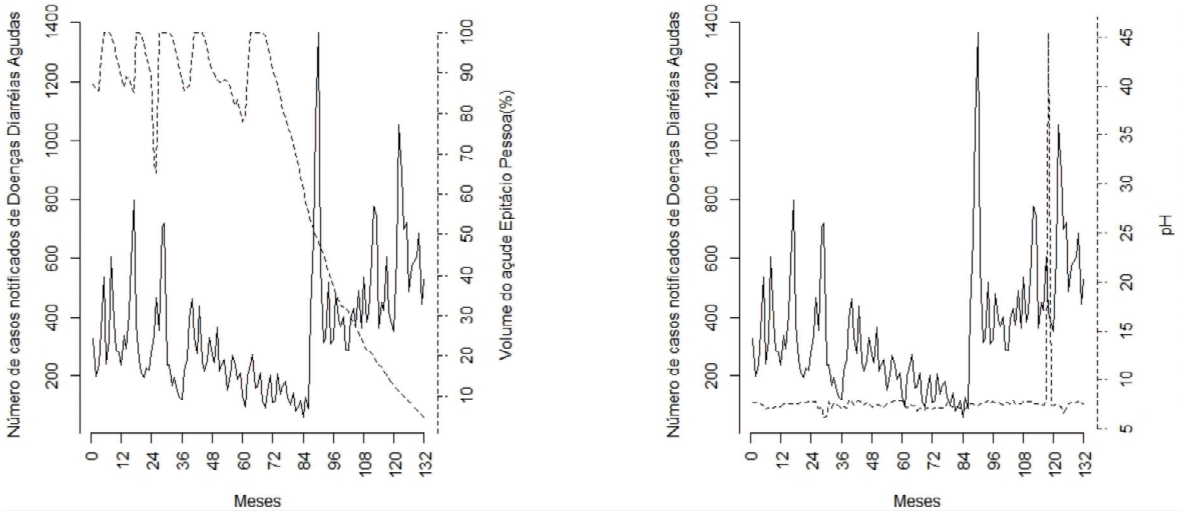
Estatísticas	**DDA**	**Volume**	**Ph**	**Cor**	**Turbidez**	**Cloro**	**Cloreto**
Média	353,735	66,53	7,77	6,94	1,38	1,525	205,775
Mediana	300,50	84,75	7,565	6,15	1,12	1,30	155,90
Variância	47898,03	1105,00	11,036	14,29	0,76	1,15	15354,62
Desvio Padrão	218,86	33,24	3,32	3,781	0,87	1,074	123,914

A posteriori foi verificado o quão relacionadas às variáveis explicativas ("Volume", "Ph", "Cor", "Turbidez", "Cloro" e "Cloreto"), coletadas no Açude Epitácio Pessoa, estavam com a variável resposta ("Número de casos notificados de crianças com DDA"). Para isto, por se tratar de dados medidos ao longo do tempo com desvios de normalidade em sua estrutura, utilizou-se a correlação de Kendall e seus respectivos valores-p.

Tabela 3 – Correlação de *Kendall* com relação a variável dependente "Número de casos notificados de crianças com Doenças Diarreicas Agudas - DDA", no município de Campina Grande, e as variáveis independentes coletadas no Açude Epitácio Pessoa.

Variáveis	**Correlação de Kendall (valor-p)**
Volume (x_1)	-0,29 (<0,001)
Ph (x_2)	0,19 (0,0021)
Cor (x_3)	-0,12 (0,0386)
Turbidez (x_4)	-0,01 (0,9169)
Cloro (x_5)	-0,15 (0,0152)
Cloreto (x_6)	0,18 (0,0020)

De acordo com a *Correlação de Kendall* (Tabela 3), pode-se observar uma relação inversamente proporcional entre a variável Volume (%) do Açude Epitácio Pessoa, com o número de casos notificados de crianças com DDA no município de Campina Grande. Isto é, à medida que o Açude foi diminuindo sua vazão, o número de crianças notificadas com DDA no município aumentou consideravelmente. Outras correlações significativas, inversamente proporcionais (valor-p<0,05), foram com as variáveis Cor e Cloro. No entanto, para as variáveis Ph e Cloreto, as correlações apresentaram significância positiva (valor-p<0,05). A relação entre essas variáveis, medidas ao longo do tempo, podem ser observadas na Figura 1.



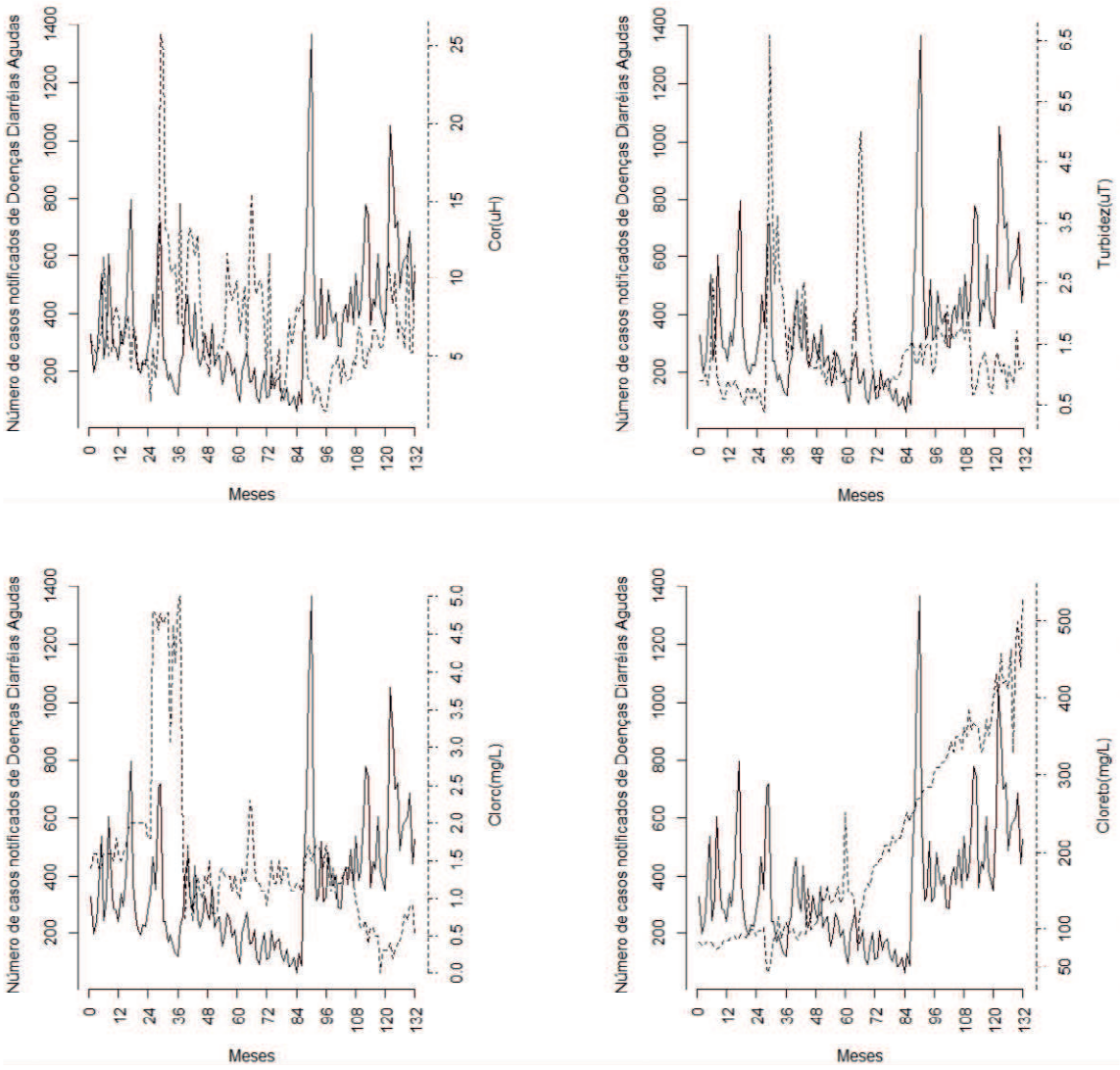


Figura 1 – Relação dos casos notificados por Doenças Diarreicas Agudas em crianças no município de Campina Grande com as variáveis Volume, Ph, Cor, Turbidez, Cloro e Cloreto, coletadas no Açude Epitácio Pessoa (Boqueirão).

Dando sequência às análises dos dados, por meio da metodologia de séries temporais, investiga-se a existência da correlação entre um valor de DDA no passado com um valor igualmente observado no presente, ou seja, a influência das notificações passadas aos eventos futuros. Analisando os gráficos da FAC e FACP (Figura 2), pode-se observar fortes indícios de que a série é não estacionária, dado que a autocorrelação decai muito lentamente para zero. Como bem observado pelo gráfico, uma vez encontrado o processo de estacionariedade, pode-se rejeitar a hipótese de não-estacionariedade da série. Porém, apenas a análise gráfica não é suficiente para evidenciar tal fato. Sendo assim, aplicou-se os testes de *Dickey-Fuller* e *Mann-Kendall* para verificar se há presença de estacionariedade e tendência, respectivamente.

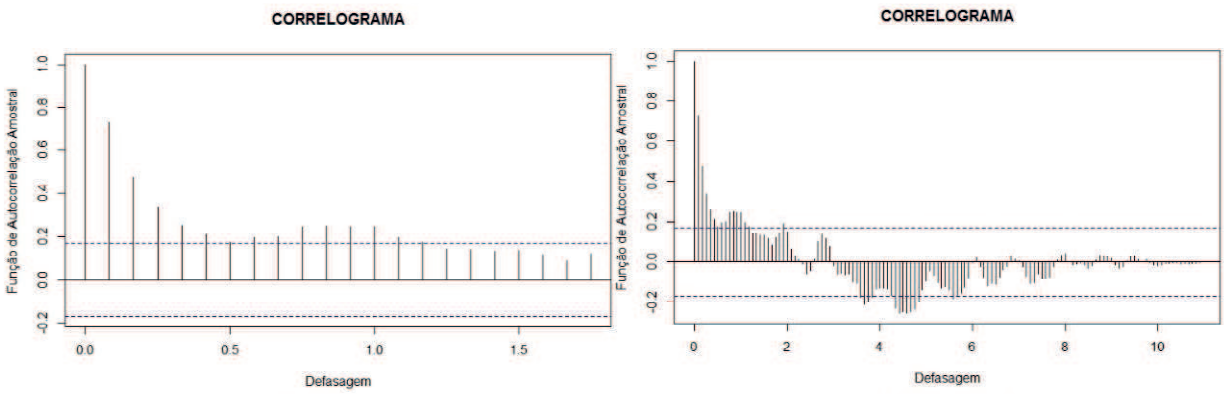


Figura 2 – Correlograma da Série Doenças Diarreicas Agudas, com *lags* iguais a 22 e 1000, respectivamente.

Avaliando a eficácia da estacionariedade de DDA pelo teste de *Dickey-Fuller*, no qual os resultados podem ser observados na Tabela 4, tem-se que a identificação do valor-p da série mensal de DDA foi de, aproximadamente, 0,05% e o valor da estatística de *Mann-Kendall* foi de 0,18. Implicando na não rejeição da hipótese nula da raiz unitária com nível de confiança de 95%, por conseguinte informando a existência de não estacionariedade na série de notificação por DDA em crianças do município de Campina Grande - PB.

Tabela 4 – Testes de *Dickey–Fuller* Aumentado e *Mann-Kendall* e seus respectivos valores-p para a série de DDA em crianças no município de Campina Grande - PB.

Variável	**ADF (valor-p)**	**Mann-Kendall (valor-p)**
DDA	-3,4429 (0,0505)	0,177 (0,002614)

3.2 Ajuste dos Modelos Poisson e Binomial Negativo com independência temporal

Primeiramente, ajustou-se o Modelo Poisson aos dados e, a partir deste ajuste todas as variáveis, com exceção da variável *dummy* "Coliformes" (qualitativa nominal), na qual a mesma foi considerada como uma variável *dummy*, foram significativas com o valor de AIC igual a 11442. Após o primeiro ajuste, a variável não significativa foi retirada do modelo e, na sequência, realizou-se um novo ajuste, em que as variáveis que permaneceram no modelo continuaram sendo significativas, apresentando um valor de AIC igual a 11441. Sendo assim, o modelo ajustado com suas respectivas estimativas foi:

$$\hat{y} = 7,85 - 0,0213x_1 - 0,0101x_2 + 0,006x_3 + 0,031x_4 + 0,02x_5 - 0,0032x_6.$$

Na sequência, foi ajustado do modelo Binomial negativo e, foi identificado, a priori, que o modelo saturado obteve um valor de AIC igual a 1706,6, com as variáveis "Volume"

e "Cloreto" significativas (valor-p<0,05); e, a partir desse resultado, foi realizado um novo ajuste com essas duas variáveis e, observou-se que o AIC diminuiu para 1699,5. Com isto, obteve-se o seguinte modelo:

$$\hat{y} = 8,5002 - 0,026x_1 - 0,005x_6.$$

Como se pode observar, o modelo mais adequado para representar os dados foi o ajuste Binomial Negativo com o modelo reduzido, segundo o critério AIC. E, isto, já era de se esperar pois, de acordo com Paula (2013), a suposição de distribuição de Poisson para a resposta é inadequada dado que a variância da variável dependente DDA foi maior que a média, resultando assim no fenômeno de *sobredispersão*. Entretanto, para concluir se de fato esse modelo foi o melhor, aplicou-se o gráfico de Envelope dos quantis teóricos com os resíduos do modelo ajustado aos dados.

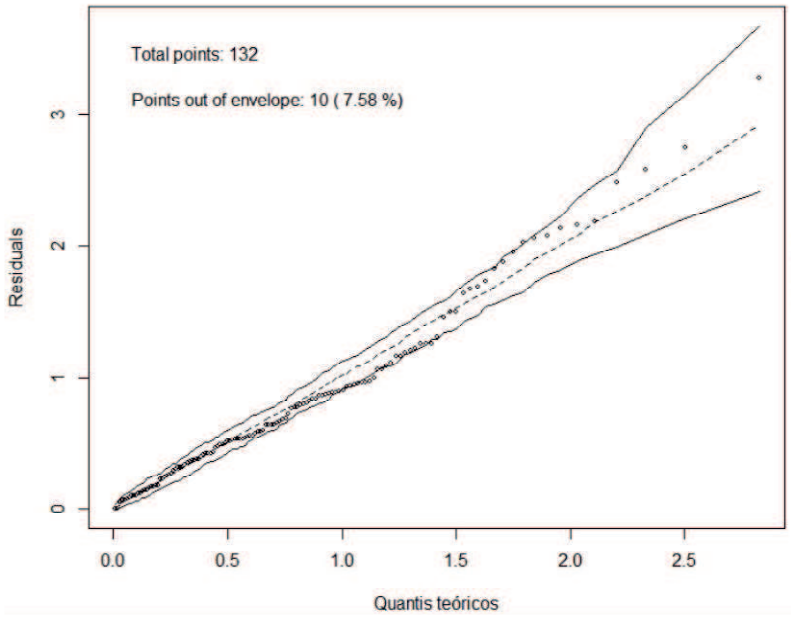


Figura 3 – Gráfico de Envelope correspondendo ao ajuste do modelo Binomial Negativo.

De acordo com a Figura 3, o modelo ainda não pareceu bem ajustado, tendo em vista que a proporção de pontos que ficaram fora do envelope de simulação foi superior a 5%. Isto é, 10 (dez) pontos ficaram fora das bandas de confiança, representando 7,58%, ainda não foi considerado como um bom ajuste, ao nível de 95% de confiança. Sendo assim, como objeto de pesquisa para este projeto, na próxima seção será apresentado novos procedimentos para a modelagem de séries temporais de contagem, que relacionam os Modelos Lineares Generalizados com estrutura autorregressiva de média móvel.

3.3 Ajuste dos Modelos Poisson e Binomial Negativo com dependência temporal

Dando sequência às análises, o modelo pertencente à série de contagem Doenças Diarreicas Agudas foi ajustado seguindo a classe dos Modelos Lineares Generalizados (MLG's). Então, para levar em conta a dependência da série temporal, foi incluído uma regressão na observação anterior. Com isto, comparou–se os ajustes das distribuições, Poisson e Binomial negativa.

Mas antes, deve–se levar em consideração as seguintes informações: o coeficiente β_1 corresponde à regressão da observação anterior; α_{13} equivale à regressão dos valores da média condicional de treze unidades (correspondente a 1 ano) de volta no tempo. Além disso, temos também duas intervenções: $interv_1$ e $interv_2$. As intervenções nada mais são do que uma quebra estrutural na série, ou seja, uma grande mudança abrupta da mesma.

Neste sentido, *outliers* ou pontos extremos também podem ser considerados como uma intervenção na série (FRANCO, 2016). E, devido a uma provável quebra na estrutura da série DDA, aplicou–se, antes de mais nada, o teste de *Chow* para a comprovação da presença ou ausência dessa quebra. Preliminarmente, encontra–se o gráfico das estatísticas F e, após isto, foi realizado o teste de *Chow*, para finalmente, comprovarmos se há ou não a presença de uma quebra estrutural.

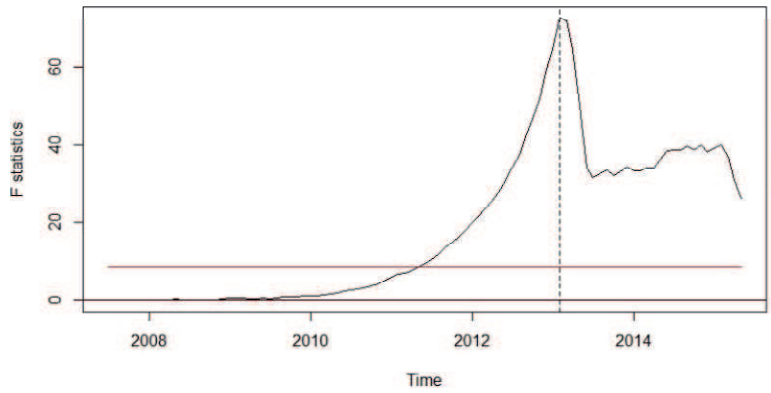


Figura 4 – Gráfico das estatísticas F com base no teste de Chow.

A Figura 4 apresenta as estatísticas F e, por meio da mesma, detectou–se uma mudança mais forte no final de 2012 para o início de 2013. Além disso, ao realizar o teste propriamente dito, observou–se que houve, de fato, uma quebra estrutural, pois o valor-p resultou em um valor menor que o nível de significância de 5% (valor–p <0,001). Diante do exposto, os resultados obtidos dos ajustes, em ambas as distribuições, encontram–se nas Tabelas 5 e 6.

Tabela 5 – Ajuste do Modelo Poisson com dependência temporal.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC(Inferior)**	**IC(Superior)**
Intercepto	74,3234	4,0868	66,3134	82,3333
β_1	0,7094	0,0103	0,68912	0,7297
α_{13}	0,0274	0,0132	0,00161	0,0532
$interv_1$	70,1654	4,7362	60,88264	79,4481
$interv_2$	0,0821	20,8455	-40,77431	40,9384

Tabela 6 – Ajuste do Modelo Binomial Negativo com dependência temporal.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC(Inferior)**	**IC(Superior)**
Intercepto	74,3234	28,7075	18,058	130,589
β_1	0,7094	0,096	0,521	0,898
α_{13}	0,0274	0,0994	-0,167	0,222
$interv_1$	70,1654	38,7013	-5,688	146,019
$interv_2$	0,0821	164,356	-322,050	322,214
σ^2	0,1498	-	-	-

Os modelos ajustados referentes às distribuições Poisson e Binomial Negativa, respectivamente, afirmam que os intervalos de confiança contêm os valores reais. E mais, em relação à Tabela 6, os resultados apresentados relatam a estimativa do coeficiente de superdispersão σ^2, que está relacionado ao parâmetro de dispersão ϕ da distribuição Binomial Negativa por $\phi = \frac{1}{\sigma^2}$. Posteriormente, foi observado por meio do critério de seleção de *Akaike* qual o ajuste representará melhor os dados.

Por meio da Tabela 7, nota–se que os modelos com dependência temporal forneceram AIC muito menores, se comparados aos critérios dos modelos com independência temporal.

Tabela 7 – Critério de AIC dos modelos Poisson e Binomial Negativo após os ajustes, com independência e dependência temporal.

Distribuição	**AIC**
Poisson	11441
Binomial Negativo	1699,5
GARMA Poisson	7092,321
GARMA Binomial Negativo	1628,119

Sendo assim, para a certificação de qual modelo dependente no tempo é, de fato, o mais adequado, foi aplicado como última ferramenta as regras de pontuação (*scoring rules*), conforme apresentado na Tabela 8.

Tabela 8 – Regras de Pontuação dos modelos dependentes no tempo e seus respectivos *scores*.

Distribuição	**Logarítmica**	**Quadrática**	**Esférica**
GARMA Poisson	26,83	0,0082	-0,0297
GARMA Binomial Negativa	6,122	-0,003	-0,0535

De acordo com a Tabela 8, levando-se em consideração as regras de pontuação, pode–se avaliar a qualidade das previsões probabilísticas, atribuindo uma pontuação numérica baseada na distribuição preditiva e no evento ou valor que se materializa. Logo, todas as regras de pontuação consideradas são favoráveis ao modelo GARMA Binomial negativo. A partir desse momento, será inicializado a análise de diagnóstico para verificar qual modelo, de fato, se ajusta melhor aos dados.

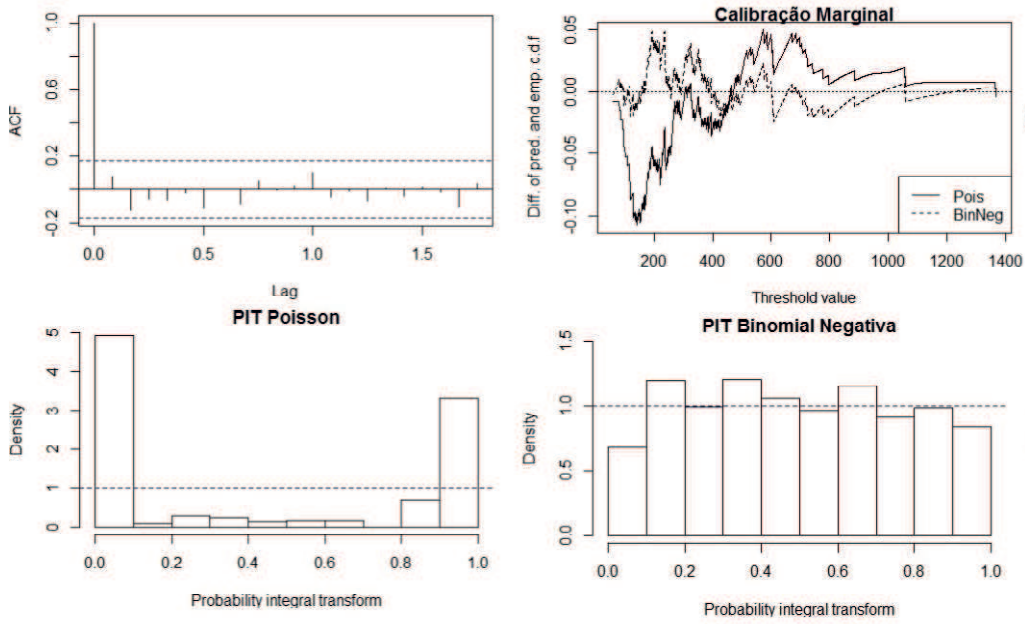


Figura 5 – Diagnósticos após os ajustes para a série de notificação de Doenças Diarreicas Agudas em crianças no município de Campina Grande.

Os resíduos de resposta são idênticos para as duas distribuições condicionais. Sua função de autocorrelação empírica, como se pode ver na Figura 5 (canto superior esquerdo), não exibe correlação serial ou sazonalidade que não foi tida em conta pelos modelos; o gráfico de calibração marginal, o qual é mostrado no canto superior direito, corrobora para o fato de que a modelo binomial negativo é o melhor dado que converge mais rapidamente para zero; já o histograma PIT (canto inferior esquerdo) em forma de U, aproximadamente, indica que a distribuição de Poisson não é adequada para o ajuste do modelo. Em oposição a isto, o histograma do PIT que corresponde à distribuição binomial negativa parece abordar melhor a uniformidade. Logo, a calibração probabilística do modelo binomial negativo é satisfatória. O grau de superdispersão não parece ser pequeno, já que o coeficiente de superdispersão estimado é de $0,1498$.

Logo, o modelo ajustado para o número de novos casos notificados de crianças com DDA, em Campina Grande, Y_t no período de tempo t é dado por $Y_t|\mathcal{F}_{t-1} \sim$

$BinNeg(\lambda_t; 6,6756)$ com

$$\lambda_t = 74,3234 + 0,7094Y_t + 0,0274\lambda_{t-13} + 70,1654\mathbb{1}(t = 93) + 0,0821\mathbb{1}(t \geq 100), \quad t = 1, ..., 132.$$

Os erros padrão dos parâmetros de regressão estimados e os intervalos de confiança correspondentes são baseados na aproximação normal. Para o coeficiente de superdispersão adicional σ^2 da distribuição binomial negativa, não existe uma aproximação analítica disponível para o seu erro padrão. Alternativamente, erros padrão (e intervalos de confiança, sendo este último, não mostrados aqui) dos parâmetros de regressão e do coeficiente de superdispersão podem ser obtidos por um *bootstrap* paramétrico.

A estimação, pelo método *bootstrap*, dos erros padrão dos parâmetros da regressão e do coeficiente de superdispersão podem ser vistas na Tabela 9, logo em seguida.

Tabela 9 – Estimação *via* Bootstrap dos Erros Padrão dos parâmetros da regressão e do coeficiente de superdispersão.

Intercepto	**beta1**	**alpha13**	**interv1**	**interv2**	**sigmasq**
48,63	0,117	0,0622	36,47	50,73	0,0217

Os erros padrão de *bootstrap* dos parâmetros de regressão são ligeiramente maiores do que àqueles baseados na aproximação normal. Notando-se também que nenhuma das abordagens reflete a incerteza adicional induzida pela seleção do modelo.

3.3.1 Ajuste dos Modelos Poisson e Binomial Negativo com a presença de variáveis explicativas

A princípio, ajustaram-se os modelos Poisson e Binomial Negativo,respectivamente, com a variável explicativa "Volume" e, então foi realizado a previsão com o modelo que melhor se adequou aos dados. Em seguida, realizou-se a mesma prática, porém, com a variável explicativa "Cloreto". Neste ajuste, obtém-se as seguintes informações: termo autorregressivo de 1º ordem (β_1); a sazonalidade por um termo autorregressivo de 12º ordem (β_{12}); a variável explicativa e a inclusão de uma covariável determinística descrevendo uma tendência linear.

Tabela 10 – Ajuste do Modelo Poisson com a presença da covariável Volume e a Tendência Linear.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC (Inferior)**	**IC (Superior)**
Intercepto	3,43835	0,287071	2,95692	4,06721
β_1	0,55789	0,033677	0,48633	0,61737
β_{12}	-0,02204	0,012354	-0,04472	0,00298
Volume	-0,00754	0,000822	-0,00932	-0,00617
Tendência Linear	-0,04097	0,006452	-0,05511	-0,02961

Tabela 11 – Ajuste do Modelo Binomial Negativo com a presença da covariável Volume e a Tendência Linear.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC (Inferior)**	**IC (Superior)**
Intercepto	3,43835	0,91041	2,2524	5,93448
β_1	0,55789	0,08003	0,3467	0,65649
β_{12}	-0,02204	0,06251	-0,1545	0,09163
Volume	-0,00754	0,00325	-0,0157	-0,00336
Tendência Linear	-0,04097	0,03107	-0,1186	0,00146
σ^2	0,14547	0,02103	0,1120	0,19335

Em ambos os ajustes, o coeficiente estimado β_1 possui um valor numérico significativo, indicando assim, uma dependência na estrutura de autocorrelação temporal do número de casos notificados de crianças com DDA no município de Campina Grande. Encontrou-se um efeito sazonal capturado pelo coeficiente de autocorrelação de décima segunda ordem β_{12}. E mais, a diminuição do Volume do Açude Epitácio Pessoa influencia o número de crianças com doenças diarreicas agudas. No entanto, o que fará decidir qual modelo ajustado escolher, será mediante os critérios de seleção apresentados na Tabela 12.

Tabela 12 – Critérios de seleção dos ajustes dos modelos GARMA Poisson e GARMA Binomial Negativo com dependência temporal.

Distribuição	**AIC**	**BIC**
GARMA Poisson	7236,513	7250,928
GARMA Binomial Negativo	1628,354	1645,651

De acordo com os resultados apresentados na Tabela 12, o modelo que melhor se ajustou aos dados em estudo foi o GARMA Binomial Negativo, dado que em ambos os critérios de seleção obteve-se valores menores. Por conseguinte, o modelo ajustado para o número de casos notificados de crianças com doenças diarreicas agudas Y_t no mês t é dado por $Y_t|\mathcal{F}_{t-1} \sim BinNeg(\lambda_t; 6,8743)$, com

$$log(\lambda_t) = 3,438 + 0,558Y_{t-1} - 0,022Y_{t-12} - 0,007X_t - 0,041t/12, \quad t = 1, ..., 132.$$

onde X_t indica o volume real do Açude Epitácio Pessoa no tempo t.

E, com base no modelo adaptado aos dados de formação até Dezembro de 2016, pode-se prever o número de casos notificados de crianças com doenças diarreicas agudas

para o mesmo ano, fazendo assim uma comparação dos valores previstos com os reais, levando-se em consideração o respectivo volume do Açude Epitácio Pessoa.

Tabela 13 – Valores observados, valores previstos e Intervalos de previsão a 95% de confiança.

Período	**Valores Observ.**	**Valores Previstos**	**IC (Inferior)**	**IC (Superior)**
Jan/2016	608	419	83	1034
Fev/2016	1055	462	88	1302
Mar/2016	885	480	81	1408
Abr/2016	698	486	75	1483
Mai/2016	724	491	84	1437
Jun/2016	487	510	85	1472
Jul/2016	574	516	89	1597
Ago/2016	587	522	84	1555
Set/2016	602	518	86	1478
Out/2016	686	524	96	1514
Nov/2016	444	530	84	1632
Dez/2016	530	535	97	1661

Conforme apresentado na Tabela 13, o modelo apresentou previsões próximas dos valores reais. Além disso, as previsões de ambos os meses constaram dentro do intervalo de 95% de confiança. Posteriormente, ajustou-se os modelos Poisson e Binomial Negativo,nesta ordem, com a variável explicativa "Cloreto" e, após a escolha do melhor ajuste, realizou–se a previsão.

Tabela 14 – Ajuste do Modelo Poisson com a presença da covariável Cloreto e a Tendência Linear.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC (Inferior)**	**IC (Superior)**
Intercepto	1,47020	0,225153	1,15694	2,02146
β_1	0,68868	0,038975	0,59083	0,74816
β_{12}	0,04878	0,03641	-0,01897	0,12319
Cloreto	0,00142	0,000146	0,00121	0,00178
Tendência Linear	-0,03534	0,004458	-0,04541	-0,02857

Tabela 15 – Ajuste do Modelo Binomial Negativo com a presença da covariável Cloreto e a Tendência Linear.

Coeficientes	**Estimativas**	**Erro Padrão**	**IC (Inferior)**	**IC (Superior)**
Intercepto	1,47020	0,58622	0,905848	3,19
β_1	0,68868	0,07276	0,484392	0,75941
β_{12}	0,04878	0,07502	-0,130435	0,16292
Cloreto	0,00142	0,00102	-0,000394	0,00353
Tendência Linear	-0,03534	0,04004	-0,110343	0,04309
sigmasq	0,15138	0,02051	0,1182	0,19335

De acordo com as Tabelas 14 e 15, os coeficientes estimados β_1 são estatisticamente significativos, indicando assim, que existe uma estrutura de autocorrelação temporal do número de casos notificados de crianças com DDA. Ademais, foi encontrado um efeito sazonal capturado pelo coeficiente de autocorrelação de décima segunda ordem β_{12}, pois o

zero está contido nos respectivos intervalos de confiança. E mais, o cloreto, presente no Açude Epitácio Pessoa, influencia no aumento do número de casos notificadas de crianças com doenças diarreicas agudas no município de Campina Grande - PB.

Tabela 16 – Critérios de seleção dos ajustes Poisson e Binomial Negativo com dependência temporal.

Distribuição	**AIC**	**BIC**
GARMA Poisson	6897,803	6912,217
GARMA Binomial Negativo	1627,316	1644,613

Com relação a Tabela 16, o modelo que melhor se ajustou aos dados em análise foi o GARMA Binomial Negativo, de acordo com os critérios de seleção. O modelo ajustado para o número de casos notificados de crianças com doenças diarreicas agudas Y_t no mês t é dado por $Y_t|\mathcal{F}_{t-1} \sim BinNeg(\lambda_t; 6,6059)$, com

$$log(\lambda_t) = 1,47 + 0,669Y_{t-1} + 0,049Y_{t-12} + 0,001X_t - 0,035t/12, \quad t = 1, ..., 132.$$

onde X_t indica o nível de cloreto real do Açude Epitácio Pessoa no tempo t.

Após a determinação do modelo, pode-se prever o número de casos notificados de crianças com doenças diarreicas agudas para o mesmo ano, fazendo assim uma comparação dos valores previstos com os reais, tendo em conta o respectivo nível de cloreto do Açude Epitácio Pessoa.

Tabela 17 – Valores observados, valores previstos e Intervalos de previsão a 95% de confiança.

Período	**Valores Observ.**	**Valores Previstos**	**IC (Inferior)**	**IC (Superior)**
Jan/2016	608	425	87	1029
Fev/2016	1055	464	70	1472
Mar/2016	885	545	74	1826
Abr/2016	698	580	81	1968
Mai/2016	724	607	62	1952
Jun/2016	487	594	76	2220
Jul/2016	574	636	79	2488
Ago/2016	587	541	69	2177
Set/2016	602	611	82	2039
Out/2016	686	674	84	2236
Nov/2016	444	657	67	2225
Dez/2016	530	736	86	2285

Como pode-se observar pela Tabela 17, o modelo apresentou previsões razoáveis com relação aos valores observados, assim como na previsão anterior, indicando assim, que as previsões foram boas para os valores selecionados.

4 Conclusão

Na abordagem em que os dados possuíam independência temporal notou-se que os valores numéricos dos critérios de seleção dos modelos se apresentaram elevados e, embora tenham preservado a significância com as variáveis Volume e Cloreto, ainda haviam problemas quanto ao ajuste.

Na abordagem em que os dados possuíam dependência temporal, observou-se que os valores numéricos dos critérios de seleção dos modelos diminuíram consideravelmente, apresentando assim, resultados mais satisfatórios que o anterior. Ao realizar as previsões foi observado que as mesmas se aproximaram dos valores observados.

Sendo assim, a classe dos modelos GARMA ajustados aos modelos de contagem de Poisson e Binomial Negativo, com inclusão de duas variáveis explicativas Volume e Cloreto, se mostraram mais eficientes que a metodologia utilizada no trabalho de Queiroz (2017). Portanto, os modelos GARMA Poisson e Binomial Negativo são, de fato, mais eficientes para ajustar dados de contagem com estrutura de autocorrelação temporal.

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 22.

ANDRADE, B. S. de. *Abordagem Estatística de Modelos para Séries Temporais de Contagem*. Dissertação (Mestrado) — Universidade Federal de São Carlos, 2013. Citado 3 vezes nas páginas 11, 18 e 27.

ANDRADE, B. S. de. *GARMA models, a new perspective using Bayesian methods and transformations*. Tese (Doutorado-Programa Interinstitucional de Pós-graduação em Estatística) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017. Citado 3 vezes nas páginas 19, 20 e 21.

ANDRADE, B. Silveira de; ANDRADE, M. G.; EHLERS, R. S. Bayesian garma models for count data. *Communications in Statistics: Case Studies, Data Analysis and Applications*, Taylor & Francis, v. 1, n. 4, p. 192–205, 2015. Citado na página 20.

BENJAMIN, M. A.; RIGBY, R. A.; STASINOPOULOS, D. M. Generalized autoregressive moving average models. *Journal of the American Statistical association*, Taylor & Francis, v. 98, n. 461, p. 214–223, 2003. Citado 3 vezes nas páginas 11, 20 e 21.

BOX, G. et al. *Time Series Analysis. Hoboken*. [S.l.]: NJ: John Wiley & Sons, Inc, 2008. Citado na página 14.

BOX, G. E.; JENKINS, G. M. Time series analysis, control, and forecasting. *San Francisco, CA: Holden Day*, v. 3226, n. 3228, p. 10, 1976. Citado 3 vezes nas páginas 13, 14 e 15.

BOX, G. E.; TIAO, G. C. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 70, n. 349, p. 70–79, 1975. Citado na página 25.

BRIËT, O. J.; AMERASINGHE, P. H.; VOUNATSOU, P. Generalized seasonal autoregressive integrated moving average models for count data with application to malaria time series with low case numbers. *PloS one*, Public Library of Science, v. 8, n. 6, p. e65761, 2013. Citado na página 20.

CHOW, G. C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 591–605, 1960. Citado na página 21.

CORDEIRO, G. M. Performance of a bartlett-type modification for the deviance. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 51, n. 2-4, p. 385–403, 1995. Citado na página 16.

CORDEIRO, G. M.; DEMÉTRIO, C. G. *Modelos lineares generalizados e Extensões*. [S.l.]: Recife: UFRPE, 2013. 483 p. Citado 3 vezes nas páginas 16, 17 e 22.

CORDEIRO, G. M.; MCCULLAGH, P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 629–643, 1991. Citado na página 18.

CZADO, C.; GNEITING, T.; HELD, L. Predictive model assessment for count data. *Biometrics*, Wiley Online Library, v. 65, n. 4, p. 1254–1261, 2009. Citado na página 26.

DICICCIO, T. J.; EFRON, B. Bootstrap confidence intervals. *Statistical science*, JSTOR, p. 189–212, 1996. Citado na página 26.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autorregressive time series with a unit root. *Journal of the American statistical association*, Taylor & Francis, v. 74, n. 366a, p. 427–431, 1979. Citado na página 24.

DICKEY, D. A.; PANTULA, S. G. Determining the order of differencing in autoregressive process. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 5, n. 4, p. 455–461, 1987. Citado na página 23.

DOBSON, A. J.; BARNETT, A. *An introduction to generalized linear models*. [S.l.]: CRC press, 2011. Citado na página 22.

FERREIRA, L. B. A. *Um estudo comparativo para modelos de séries temporais de contagem*. Dissertação (Mestrado) — Instituto de Ciências Exatas da Universidade Federal de Minas Gerais. Departamento de Estatística, 2015. Citado 3 vezes nas páginas 11, 13 e 15.

FOKIANOS, K.; FRIED, R. Interventions in ingarch processes. *Journal of Time Series Analysis*, Wiley Online Library, v. 31, n. 3, p. 210–225, 2010. Citado na página 25.

FOKIANOS, K.; FRIED, R. Interventions in log-linear processes. *Statistical Modelling*, Sage Publications Sage India: New Delhi, India, v. 12, n. 4, p. 299–322, 2012. Citado na página 25.

FRANCO, G. da C. Apostila de modelos lineares em séries temporais. 2016. Citado na página 33.

GNEITING, T.; BALABDAOUI, F.; RAFTERY, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 69, n. 2, p. 243–268, 2007. Citado 2 vezes nas páginas 25 e 26.

HIPEL, K. W.; MCLEOD, A. I. *Time series modelling of water resources and environmental systems*. [S.l.]: Elsevier, 1994. v. 45. Citado na página 24.

MANN, H. B. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 245–259, 1945. Citado na página 25.

MILHORANÇA, I. A. *Modelos Paramétricos para Séries Temporais de Contagem*. Dissertação (Mestrado) — Instituto de Matemática e Estatística, Universidade de São Paulo, 2014. Citado na página 15.

MORETTIN, P. A.; TOLOI, C. *Análise de séries temporais*. [S.l.]: Blucher, 2006. Citado 2 vezes nas páginas 13 e 14.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society*, Wiley Online Library, v. 135, p. 370–384, 1972. Citado 3 vezes nas páginas 11, 15 e 17.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. Contato: giapaula@ime.usp.br: Instituto de Matemática e Estatística da Universidade de São Paulo, 2013. 440 p. Citado 2 vezes nas páginas 11 e 32.

QUEIROZ, L. F. R. *Doenças Diarreicas Agudas: ocorrência em crianças com idade entre zero a nove anos e a correlação com a qualidade da água*. Dissertação (Mestrado em Recursos Naturais) — Universidade Federal de Campina Grande, 2017. Citado 2 vezes nas páginas 28 e 41.

RESENDE, M. D. V.; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. *Rev. Mat. Estat*, v. 20, p. 39–65, 2002. Citado na página 16.

SANTOS, D. A. da S. et al. Redução de infecção respiratória aguda em crianças menores de dois anos em rondonópolis-mt. *Revista de Epidemiologia e Controle de Infecção*, v. 7, n. 1, 2017. Citado na página 11.

SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, v. 6, n. 2, p. 461–464, 1978. Citado na página 23.

STASINOPOULOS, M. D. et al. *Flexible Regression and Smoothing: Using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2017. v. 1. Citado na página 18.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. Regression models for count data in r. Department of Statistics and Mathematics, WU Vienna University of Economics and Business, 2007. Citado na página 11.