Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Graduação em Estatística

Comparação de Desempenho e Adequação de três processos para séries temporais INAR

Paulo Manoel da Silva Junior

João Pessoa - PB
2025

Paulo Manoel da Silva Junior

Comparação de Desempenho e Adequação de três processos para séries temporais INAR

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Estatística.

Orientador: Profª. Drª. Tatiene Correia de Souza

João Pessoa - PB
2025

Agradecimentos

Em algum momento da minha vida parei de prestar atenção no presente e passei a viver na ansiedade de um futuro melhor. "Preocupado com uma única folha, você não verá a árvore. Preocupado com uma única árvore você não perceberá toda a floresta. Não se preocupe com um único ponto. Veja tudo em sua plenitude sem se forçar."– Takuan Sōhō. Obrigado Eraldo do passado por decidir mudar.

Meu maior agradecimento é pra mainha e painho, Dona Adeilda e Seu Everaldo. Mãe obrigado por todo o carinho infindável que a senhora me proporciona. Pai obrigado por trabalhar tão duro por nós. Obrigado aos dois pelos valores ensinados, isso nenhuma escola ou universidade poderia me ensinar.

Meu sincero obrigado, em especial, ao professor Marcelo por aceitar me orientar nessa etapa final da graduação. Ao professor Tablada e a professora Everlane, pelo suporte durante minha participação na tutoria e na pesquisa científica, projetos esses na qual só agregaram positivamente na minha vida e carreira. Professora Gilmara, obrigado pela paciência, por toda educação e prestatividade como coordenadora desse curso. Por fim, e não menos importante a todos os professores na qual eu pude ter a oportunidade de ser aluno.

Paulin, Eltin, Arthur, Gleyce, nossos dias resolvendo listas e mais listas de exercícios valeram a pena. A troca de conhecimento que tivemos durante essa jornada foi fundamental na formação da pessoa que sou hoje.

Resumo

Este trabalho aborda o uso da aprendizagem de máquina na tentativa de prever o próximo dia de compra de clientes a partir do histórico de compras. Os dados foram obtidos de uma distribuidora da Paraíba que fornece uma vasta variedade de produtos de diferentes departamentos, descartáveis, químicos, papelarias, equipamentos de proteção. Diante do cenário competitivo e da explosão de dados no mercado, a inteligência comercial tornou-se essencial para antecipar desafios e identificar oportunidades. O estudo abrange a coleta e análise do histórico de compras, além da utilização de segmentação de clientes. Diferentes algoritmos, como Redes Neurais Artificiais, Máquinas de Vetores de Suporte e Florestas Aleatórias, são avaliados em busca dos melhores resultados. Os objetivos incluem extrair informações relevantes, realizar análise exploratória, segmentar clientes e avaliação de modelos e técnicas do campo da aprendizagem de máquina. Espera-se que os resultados contribuam para personalizar estratégias do setor comercial.

Palavras-chave: Aprendizagem de máquina, Inteligência de Mercado, Segmentação RFV, Regressão.

Abstract

This study explores the application of machine learning techniques to predict the next purchase day of customers based on their purchase history. The data was obtained from a company distributor in Paraíba, that provides a wide variety of products from different departments, including disposables, chemicals, stationery, and protective equipment. Given the competitive landscape and the data explosion in the market, business intelligence has become crucial for anticipating challenges and identifying opportunities. The research encompasses the collection and analysis of purchase histories, as well as the utilization of customer segmentation. Various algorithms, such as Artificial Neural Networks, Support Vector Machines, and Random Forests, are evaluated to attain optimal results. Objectives include extracting relevant information, conducting exploratory analysis, segmenting customers, and assessing models and techniques within the field of machine learning. The anticipated outcome is to enhance the personalization of commercial strategies.

Keywords: Machine learning, Business Intelligence, RFV Segmentation, Regression.

Lista de tabelas

Lista de ilustrações

Sumário

1 Introdução

As séries temporais com dados inteiros não negativos, conhecidas como **séries de contagem**, ou **pontos de contagem**, são comuns em diversas áreas, como finanças, epidemiologia, seguros e telecomunicações. Exemplos típicos incluem o número de chamadas telefônicas por hora, a quantidade de novos clientes inadimplentes por dia, a quantidade de seguros vendidos a cada semana, acidentes por dia ou casos de doenças por semana. Por sua natureza discreta, essas séries desafiam a aplicação direta de modelos clássicos como os AR, MA, ARMA ou ARIMA, os quais pressupõem variáveis contínuas e normalmente distribuídas (CAMERON; TRIVEDI, 1998).

Como alternativa aos modelos tradicionais, surgiram os chamados **modelos autorregressivos para dados inteiros não negativos**, notadamente os processos **INAR (Integer value autorregresive)**, que respeitam a estrutura discreta da variável ao incorporar um operador de redução. A formulação original do processo INAR(1) foi introduzida inicialmente por Al-Osh e Alzaid (1987) e McKenzie (1985), sendo baseado no operador *thinning binomial* proposto por steutel1979, o qual permite definir processos autorregressivos respeitando a natureza discreta da variável.

Modelos INAR são particularmente atrativos por possibilitarem a incorporação de dependência temporal entre observações, algo que é essencial em séries de contagem. No modelo INAR(1), essa dependência é representada por meio do operador *thinning binomial*, denotado por $\alpha \circ X_{t-1}$, onde $\alpha \in [0, 1]$ representa a taxa de retenção da série. O termo de inovação, denotado por ϵ_t, por outro lado, representa a contribuição probabilistica que é adicionada ao processo a cada instante do tempo. Geralmente, ele é modelado por uma distribuição discreta, sendo a distribuição Poisson a mais utilizada. Essa escolha facilita as análises e resultados, mas a escolha dessa distribuição traz uma limitação que deve ser levada em consideração, pois, ela pressupõe que a média e variância são iguais. Essa pressuposição em muitos casos não é verdadeira, pois, é observado em muitos casos que a variabilidade é maior ou menor do que a média McKenzie (2003).

Nas últimas décadas, os processos INAR têm se mostrado úteis em diversas aplicações práticas. (PEDELI; KARLIS, 2011) propuseram uma versão bivariada do modelo INAR(1) para descrever o número diário de transações financeiras em bolsas de valores, considerando a correlação entre pares de séries de contagem. Em outro contexto, (FOKIANOS; TJøSTHEIM, 2009) utilizaram modelos baseados em INAR para modelar a frequência de chamados de emergência em serviços hospitalares, com foco na dependência temporal e na sobredispersão dos dados. Já (ZHANG; WANG; ZHU, 2019) aplicaram modelos INAR para analisar séries mensais de casos de tuberculose na China, evidenciando a adequação do modelo em contextos epidemiológicos. Tais aplicações reforçam a relevância da modelagem de séries de contagem com sobredispersão e motivam a utilização de distribuições alternativas à Poisson, tais como a binomial negativa e a geométrica.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo analisar o desempenho de dois estimadores dos parâmetros do processo Poisson, Binomial Negativa e Geométrica INAR(1), sendo os estimadores mínimos quadrados condicionais e máxima verossimilhança condicional. Como um segundo objetivo desejamos propor e comparar os mesmos dados as três distribuições consideradas, usando os métodos de estimação e depois realizar previsões um passo a frente. Os estudos de desempenho são feitos por meio de simulações de Monte Carlo, que serão realizadas usando a linguagem R (R, Core Team, 2025).

Para alcançar esse objetivo geral, estabelecem-se os seguintes objetivos específicos:

1.1.2 Objetivos Específicos

1. Conduzir estudos de simulação, variando cenários de parâmetros, a fim de comparar os estimadores dos modelos analisados;

2. Avaliar o desempenho dos modelos utilizando medidas de qualidade de ajuste;

3. Aplicar os modelos a dados reais de uma seguradora;

4. Comparar os resultados dos modelos estatísticos com os métodos de estimação propostos.

2 Referencial Teórico

2.1 Processo INAR

A relevância dos modelos INAR pode ser destacada a partir de algumas de suas principais características:

- **Respeito à natureza discreta dos dados:** O modelo em questão assegura que tanto os valores ajustados, quanto as previsões permaneçam dentro do conjunto dos inteiros não negativos, refletindo adequadamente a essência dos dados de contagem.
- **Depêndencia temporal adequada:** Diferentemente dos modelos lineares clássicos, os processos INAR descrevem a correlação entre observações sucessivas por meio do operador de *thinning*, baseado na convolução binomial. Esse mecanismo representa de maneira mais realista a influência de eventos passados sobre as contagens futuras.
- **Flexibilidade na escolha da distribuição:** Os termos de inovação do modelo podem seguir diferentes distribuições, como Poisson, binomial negativa ou geométrica, o que amplia a sua capacidade de lidar com padrões de dispersão e variabilidade frequentemente observados em séries de contagem.
- **Interpretação intuitiva dos parâmetros:** O coeficiente autorregressivo pode ser entendido como a fração média de eventos do período anterior que permanece no presente, enquanto o termo de inovação capta a ocorrência de novos eventos independentes.
- **Base teórica consolidada:** Os processos INAR contam com uma formulação estatística bem estruturada, contemplando condições de estacionariedade, funções de autocorrelação e resultados assintóticos para estimadores.

2.1.1 Operador *thinning* binomial

Uma séria temporal pode ser definida como: Seja $Y_t, t \in Z$ uma série temporal. Dizemos que $\{Y_t, t \in Z\}$ é um processo autoregressivo, AR(1) se satisfaz a equação recursiva

$$Y_t = \alpha Y_{t-1} + \epsilon_t$$

Onde $\alpha \in R$ e a sequência $\{\epsilon_t, t \in R\}$ chamada de ruído branco, é uma coleção de variáveis aleatórias não correlacionadas, a média e variância dependem da distribuição.

Como em muitos casos $\alpha \in R$, para uma série temporal de valores inteiros não negativos não é admitido esta representação, mesmo que seja considerado uma distribuição de valores inteiros não negativos para ϵ_t, Y_t não será, necessariamente um valor inteiro.

Na literatura uma das formas de obter modelos que garantem que Y_t seja um valor inteiro não negativo é utilizar um operador conveniente para garantir que as variáveis aleatórias assumam valores inteiros não negativos. Utilizaremos o operador *thinning* binomial, proposto por (STEUTEL; HARN, 1979).

Definição 1: Seja Y uma variável aleatória não negativa que assume valores inteiros e seja $\alpha \in [0, 1]$. O operador *thinning* binomial é definido como:

$$\alpha \circ Y = \sum_{i=1}^{Y} N_i,$$

Em que as variáveis aleatórias $N_i^{\prime s}$ são independentes e identicamente distribuídas (i.i.d.), com distribuição Bernoulli de parâmetro α. As variáveis $N_i^{\prime s}$ são chamadas de séries de contagem de $\alpha \circ Y$.

2.1.1.1 Propriedades do operador *thinning* binomial

Sejam X e Y variáveis aleatórias não negativas assumindo valores inteiros. Sejam α e β constantes que pertencem aos reais no intervalo $[0, 1]$ e suponha que a série de contagem de $\alpha \circ Y$ é independente da série de contagem de $\beta \circ X$ são independentes de X e de Y. Então:

1. $0 \circ Y = 0$
2. $1 \circ Y = Y$
3. $\alpha \circ (\beta \circ Y) \stackrel{d}{=} (\alpha\beta) \circ Y$
4. $\alpha \circ (Y + X) \stackrel{d}{=} (\alpha \circ Y) + (\beta \circ X)$
5. $E(\alpha \circ Y) = \alpha E(Y)$
6. $Var(\alpha \circ Y) = \alpha^2 Var(Y) + \alpha(1 - \alpha)E(Y)$

em que $X \stackrel{d}{=} Y$ significa que X e Y têm a mesma distribuição.

As demonstrações destas propriedades, bem como de outras do operador *thinning* binomial já foram demonstradas e podem ser encontradas em Gomes (2009), Barcelos (2008) e Silva (2005).

2.1.2 Processo INAR(1)

A partir do que foi definido sobre o operador *thinning* binomial, é possível definir o processo autoregressivo de valores inteiros de ordem um, INAR(1), que foi proposto por (AL-OSH; ALZAID, 1987) e (MCKENZIE, 1985).

Definição 2: Um processo estocástico discreto de valores inteiros não negativos $\{Y_t, t \in Z\}$ é dito ser um processo INAR (1), se satisfaz a seguinte equação de recursão

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t, \tag{2.1}$$

em que $\alpha \in (0, 1)$, $\{\epsilon_t, t \in Z\}$ é uma sequência de variáveis aleatórias i.i.d. de valores inteiros não negativos tal que $E(\epsilon_t) = \mu_\epsilon$, $Var(\epsilon_t) = \sigma_\epsilon^2$ e ϵ_t é independente de Y_s para $s < t$.

O processo INAR (1) pode ser interpretado da seguinte forma: Y_t representa a população de habitantes remanescentes em uma região no caso de migração no tempo t. $\alpha \circ Y_{t-1}$ representa os habitantes daquela região que permaneceram naquela região no tempo $t - 1$, ϵ_t representa os novos haabitantes que chegaram para morar naquela região no tempo t e $Y_{t-1} - \alpha \circ Y_{t-1}$ a quantidade de habitantes que saíram daquela região.

De acordo com Du e Li (1991) e Latour (1998) se $\alpha < 1$. então o processo INAR(1) é estacionário. Consideraremos aqui apenas quando $\alpha < 1$, caracterizando assim um processo estacionário.

Segue algumas propriedades do processo INAR(1) que segue para adequação de acordo com a distribuição de ϵ_t

2.1.2.1 Propriedades do Processo INAR(1)

Seja $\{Y_t, t \in Z\}$ um processo INAR(1), com $\mu_\epsilon = E(\epsilon_t)$ e $\sigma_\epsilon^2 = Var(\epsilon_t)$. Então, $\{Y_t, t \in Z\}$ satisfaz as seguintes propriedades:

1. $E(Y_t) = \frac{\mu_\epsilon}{1-\alpha}$
2. $Var(Y_t) = \frac{\sigma_\epsilon^2 + \alpha\mu_\epsilon}{1-\alpha^2}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu_\epsilon$
4. $Var(Y_t|Y_{t-1}) = \alpha(1-\alpha)Y_{t-1} + \sigma_\epsilon^2$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j Var(Y_t)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.2 Distribuições associadas

2.2.1 Processo Poisson INAR(1)

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou seja, um processo em que Y_t satisfaz a equação 2.1.

Quando $\{\epsilon_t, t \in Z\}$ é um conjunto de variáveis aleatórias independentes com distribuição Poisson de parâmetro λ, então $\{Y_t, t \in Z\}$ é chamado de processo Poisson INAR(1). E nesse caso, a média e variância são iguais. Ou seja, $\mu_\epsilon = \sigma_\epsilon^2 = \lambda$.

Seja $\{Y_t, t \in Z\}$ um processo Poisson INAR(1) então, considerando as propriedades do operador *thinning* binomial que foram apresentadas na seção 2.1.2.1 e também que temos um processo estacionário, ou seja, a média e variância permanecem constantes ao longo do tempo. É possível encontrar as seguintes propriedades.

1. $E(Y_t) = \frac{\lambda}{1-\alpha}$
2. $Var(Y_t) = \frac{\lambda}{1-\alpha}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \lambda$
4. $Var(Y_t|Y_{t-1}) = \alpha(1-\alpha)Y_{t-1} + \lambda$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j Var(Y_t)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.2.2 Processo Geométrica INAR(1)

Conforme sugerido por (MCKENZIE, 1986) em processos INAR(1) podem ser utilizados a distribuição **Geométrica e Binomial Negativa**, sendo assim.

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou seja, um processo em que Y_t satisfaz a equação (2.1). Quando $\epsilon_t, t \in Z$ é um conjunto de variáveis aleatórias independentes com distribuição Geométrica parametrizada pela média, então $\{Y_t, t \in Z\}$ é chamado de processo Geométrica INAR(1). E nesse caso, temos os seguintes resultados:

Para uma distribuição geométrica com em 0,1,2,3,...

$$P(X = k) = (1-p)^k p, \quad k = 0, 1, 2, 3, ...$$

Parametrizando pela média, temos os seguintes resultados.

- $p = \frac{1}{1+\mu}$
- $P(X = k) = \frac{\mu^k}{(\mu+1)^{k+1}}$

- $E(X) = \mu$
- $Var(X) = \frac{\mu}{(\mu+1)^2}$

A distribuição aqui foi parametrizada pela média e $k \in 0, 1, 2, 3, ...$ As propriedades do processo seguem abaixo:

1. $E(Y_t) = \frac{\mu}{1-\alpha}$
2. $Var(Y_t) = \frac{\alpha\mu + \frac{\mu}{(\mu+1)^2}}{(1-\alpha)^2}$
3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu$
4. $Var(Y_t|Y_{t-1}) = Y_{t-1}\alpha(1-\alpha) + \frac{\mu}{(\mu+1)^2}$
5. $Cov(Y_t, Y_{t+j}) = \alpha^j \left(\frac{\alpha\mu + \frac{\mu}{(\mu+1)^2}}{(1-\alpha^2} \right)$, para $j \in N$
6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h$, para $h \in N$

2.2.3 Processo Binomial Negativa INAR(1)

Considere $\{Y_t, t \in Z\}$ um processo INAR(1), ou sejam um processo em que Y_t satisfaz a equação (2.1). Quando $\epsilon_t, t \in Z$ é um conjunto de variáveis aleatórias independentes com distribuição Binomial Negativa parametrizada pela média, então $\{Y_t, t \in Z\}$ é chamado de processo Binomial Negativa INAR(1). Temos os seguintes resultados, considerando uma distribuição parametrizada pela média:

$$p = \frac{\mu}{\sigma^2}$$

$$P(x = k) = \binom{k + \frac{\mu^2}{\sigma^2-\mu} - 1}{k} \left(1 - \frac{\mu}{\sigma^2}\right)^k \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}$$

$$r = \frac{\mu^2}{\sigma^2 - \mu}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

A distribuição foi parametrizada pela média, e as propriedades do processo segue abaixo:

1. $E(Y_t) = \frac{\mu}{1-\alpha}$

2. $Var(Y_t) = \frac{\alpha\mu+\sigma^2}{1-\alpha^2}$

3. $E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \mu$

4. $Var(Y_t|Y_{t-1}) = y_{t-1}\alpha(1-\alpha) + \sigma^2$

5. $Cov(Y_t, Y_{t+j}) = \alpha^j \left(\frac{\alpha\mu+\sigma^2}{1-\alpha^2}\right), \quad \text{para } j \in N$

6. $\rho_Y(h) = Corr(Y_t, Y_{t+h}) = \alpha^h, \quad \text{para } h \in N$

2.3 Métodos de Estimação

2.3.1 Estimadores de Mínimos Quadrados Condicionais

Seja $Y_1, Y_2, Y_3, ..., Y_n$ uma amostra do processo INAR(1) dado na equação (2.1). Estamos interessados em estimar o vetor de parâmetros $\theta = (\alpha, \lambda)$. Sabemos que:

$$E(Y_t|Y_{t-1}) = \alpha Y_{t-1} + \lambda = g(\theta, Y_{t-1})$$

Considere a função,

$$Q_n(\theta) = \sum_{t=2}^{n}[Y_t - g(\theta, Y_{t-1})]^2$$

Os estimadores de MQC de α e λ são os valores de α e λ que minimizam $Q_n(\theta)$. Depois de derivar $Q_n(\theta)$ em relação a α e λ e sendo igualhada as duas derivadas a zero, obtemos que os estimadores de Minímos quadrados condicionais de α e λ são dados por:

$$\widehat{\alpha}_{MQC} = \frac{\sum_{t=2}^{n} Y_t Y_{t-1} - \frac{1}{n-1}\sum_{t=2}^{n} Y_t \sum_{t=2}^{n} Y_{t-1}}{\sum_{t=2}^{n} Y_{t-1}^{n} Y_{t-1}^2 - \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_{t-1}\right)^2} \tag{2.2}$$

$$\widehat{\lambda}_{MQC} = \frac{1}{n-1}\left(\sum_{t=2}^{n} Y_t - \widehat{\alpha}_{MQC}\sum_{t=2}^{n} Y_{t-1}\right) \tag{2.3}$$

2.3.2 Estimadores de Máxima Verossimilhança Condicional

Para o estimador de máxima verossimilhança condicional, a distribuição deve ser levada em consideração, por isso, teremos funções diferentes, a depender da Distribuição Associada.

1. **Processo Poisson**

   Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Poisson (λ_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α). Logo, $\alpha \circ Y \sim$ Binomial (γ, α)

A função de log-verossimilhança é dada por:

$$l(\alpha, \lambda) = \sum_{t=2}^{T} ln \;\; P(Y_t|Y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$\begin{aligned} P(Y_t = y_t|Y_{t-1} = y_{t-1}) &= P(\alpha \circ Y_{t-1} + \epsilon_t|Y_{t-1} = y_{t-1}) \\ &= P(\alpha \circ Y_{t-1} + \epsilon_t = y_t) \end{aligned}$$

Em que $\alpha \circ Y_{t-1} \sim$ Binomial (y_{t-1}, α) e $\epsilon_t \sim$ Poisson (λ), com $X \perp Y$, então,

$$\begin{aligned} P(X + Y) &= \sum_{j=0}^{k} P(X = j, Y = k - j) \\ &= \sum_{j=0}^{k} P(X = j)P(Y = k - j) \end{aligned}$$

Note que, $0 \leq j \leq n$ e $0 \leq j \leq k$. Logo, $0 \leq j \leq min(n.k)$

Logo,

$$\begin{aligned} P(X + Y = k) &= \sum_{j=0}^{min(n,k)} = P(x = j)P(Y = k - j) \\ &= \sum_{j=0}^{min(n,k)} \binom{n}{j} \alpha^j (1 - \alpha)^{n-j} \frac{e^{-\lambda}\lambda^{k-j}}{(k - j)!} \end{aligned}$$

Portanto, a função de log-verossimilhança de um processo poisson INAR (1) é dado por:

$$P(Y_t = y_t|Y_{t-1} = y_{t-1}) = \sum_{j=0}^{min(y_{t-1},y_t)} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{e^{-\lambda_t}\lambda_t^{y_t-j}}{(y_t - j)!}$$

$$l(\alpha, \lambda) = \sum_{t=2}^{T} ln \left(\sum_{j=0}^{Min(y_t,y_{t-1})} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{e^{-\lambda t}\lambda t^{y_t-j}}{(y_t - j)!} \right)$$

2. **Processo Geométrica**

Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Geométrica (k_t, μ_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α).

A função de log-verossimilhança é dada por:

$$l(\alpha, \mu) = \sum_{t=2}^{T} ln \;\; P(Y_t = y_t | Y_{t-1} = y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$\begin{aligned} P(Y_t = y_t | Y_{t-1} = y_{t-1}) &= P(\alpha \circ Y_{t-1} + \epsilon_t | Y_{t-1} = y_{t-1}) \\ &= P(\alpha \circ Y_{t-1} + \epsilon_t = y_t) \end{aligned}$$

Em que $\alpha \circ Y_{t-1} \sim$ Binomial (y_{t-1}, α) e $\epsilon_t \sim$ Geométrica (k, μ), com $X \perp Y$, então,

$$\begin{aligned} P(X + Y) &= \sum_{j=0}^{k} P(X = j, Y = k - j) \\ &= \sum_{j=0}^{k} P(X = j) P(Y = k - j) \end{aligned}$$

Note que, $0 \leq j \leq n$ e $0 \leq j \leq k$. Logo, $0 \leq j \leq min(n.k)$

Logo,

$$\begin{aligned} P(X + Y = k) &= \sum_{j=0}^{min(n,k)} = P(x = j) P(Y = k - j) \\ &= \sum_{j=0}^{min(n,k)} \binom{n}{j} \alpha^j (1 - \alpha)^{n-j} \frac{\mu^{k-j}}{(\mu + 1)^{k-j+1}} \end{aligned}$$

Portanto, a função de log-verossimilhança de um processo geométrica INAR (1) parametrizado por μ é dado por:

$$P(Y_t = y_t | Y_{t-1} = y_{t-1}) = \sum_{j=0}^{min(y_{t-1}, y_t)} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{\mu^{y_t - j}}{(\mu + 1)^{y_t - j + 1}}$$

$$l(\alpha, \mu) = \sum_{t=2}^{T} ln \left(\sum_{j=0}^{Min(y_t; y_{t-1})} \binom{y_{t-1}}{j} \alpha^j (1 - \alpha)^{y_{t-1}-j} \frac{\mu^{y_t - j}}{(\mu + 1)^{y_t - j + 1}} \right)$$

3. **Processo Binomial Negativa**

Para o estimador de máxima verossimilhança condicional, temos que $Y_t = \alpha \circ Y_{t-1} + \epsilon_t$, em que $\epsilon_t \sim$ Bin. negativa (r_t, P_t)

O operador *thinning binomial*: $\alpha \circ Y = \sum_{K=0}^{y} B_K, \quad B_K \sim$ Bernoulli (α).

A função de log-verosssimilhança é dada por:

$$l(\alpha,\mu,\sigma^2)=\sum_{t=2}^{T} ln\ \ P(Y_t=y_t|Y_{t-1}=y_{t-1})$$

A probabilidade de Transição: $P(y_t|y_{t-1})$ é:

$$\begin{aligned}P(Y_t=y_t|Y_{t-1}=y_{t-1}) &= P(\alpha\circ Y_{t-1}+\epsilon_t|Y_{t-1}=y_{t-1})\\ &= P(\alpha\circ Y_{t-1}+\epsilon_t=y_t)\end{aligned}$$

Em que $\alpha\circ Y_{t-1}\sim$ Binomial (y_{t-1},α) e $\epsilon_t\sim$ Bin. Negativa (r,p), com $X\perp Y$, então,

$$\begin{aligned}P(X+Y) &= \sum_{j=0}^{k} P(X=j,Y=k-j)\\ &= \sum_{j=0}^{k} P(X=j)P(Y=k-j)\end{aligned}$$

Note que, $0\leq j\leq n$ e $0\leq j\leq k$. Logo, $0\leq j\leq min(n.k)$

Logo,

$$\begin{aligned}P(X+Y=k) &= \sum_{j=0}^{min(n,k)} = P(x=j)P(Y=k-j)\\ &= \sum_{j=0}^{min(n,k)}\binom{n}{j}\alpha^j(1-\alpha)^{n-j}\binom{k-j+\frac{\mu^2}{\sigma^2-\mu}-1}{k-j}\left(1-\frac{\mu}{\sigma^2}\right)^{k-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}\end{aligned}$$

Portanto, a função de log-verossimilhança de um processo INAR (1) com distribuição binomial negativa, parametrizada por μ e σ^2 é dada por:

$$\begin{aligned}P(Y_t=y_t|Y_{t-1}=y_{t-1}) = &\sum_{j=0}^{min(y_{t-1},y_t)}\binom{y_{t-1}}{j}\alpha^j(1-\alpha)^{y_{t-1}-j}\binom{y_t-j+\frac{\mu^2}{\sigma^2-\mu}-1}{y_t-j}\\ &\left(1-\frac{\mu}{\sigma^2}\right)^{y_t-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}\end{aligned}$$

$$\begin{aligned}P(Y_t=y_t|Y_{t-1}=y_{t-1}) = &\sum_{j=0}^{min(y_{t-1},y_t)}\binom{y_{t-1}}{j}\alpha^j(1-\alpha)^{y_{t-1}-j}\binom{y_t-j+\frac{\mu^2}{\sigma^2-\mu}-1}{y_t-j}\\ &\left(1-\frac{\mu}{\sigma^2}\right)^{y_t-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}\end{aligned}$$

A função final de log-verossimilhança de um processo com distribuição binomial negativa é dada por:

$$l(\alpha,\mu,\sigma^2) = \sum_{t=2}^{T}\left(\sum_{j=0}^{Min(y_t,y_{t-1})}\binom{y_{t-1}}{j}\alpha^j(1-\alpha)^{y_{t-1}-j}\binom{y_t - j + \frac{\mu^2}{\sigma^2-\mu} - 1}{y_t - j}\right.$$
$$\left.\left(1-\frac{\mu}{\sigma^2}\right)^{y_t-j}\left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{(\sigma^2-\mu)}}\right)$$

2.4 Previsão do Processo INAR(1)

Considere a série $Y_1, \cdots, Y_t$ uma amostra do processo INAR (1). Tomando como verdadeiro a premissa de que conhecemos a série até o tempo t, estamos interessados em fazer a previsão de Y_{t+1}.

O valor de k que minimiza o Erro Quadrático Médio Condicional (EQMC), dado por,

$$E\left[(Y_{t+1}-k)^2\,|Y_t\right]$$

é a esperança de Y_{t+1} dado Y_t, ou seja, $k = E\left[Y_{t+1}|Y_t\right]$, e usaremos para fazer a previsão de Y_{t+1}.

O valor de a que minimiza o Erro Absoluto Médio Condicional (EAMC), dado por,

$$E = [|Y_{t+1}-a||Y_t]$$

corresponde à mediana da distribuição condicional de Y_{t+1} dado Y_t. Então, usaremos a mediana de Y_{t+1} dado Y_t como previsão do valor de Y_{t+1}.

2.4.1 Média da distribuição condicional um passo a frente

Como foi observado na seção 2.4, a esperança condicional de Y_{t+1} dado Y_t minimiza o EQMC. Assim, podemos usar esta esperança como previsão de Y_{t+k} dado que conhecemos a série até Y_t. Em sua tese de Doutorado, Freeland (1998), demonstrou que:

$$E\left(Y_{t+k}|Y_t\right) = \alpha^k Y_t + \mu_\epsilon \frac{1-\alpha^k}{1-\alpha}$$

Substituindo k por 1, temos que:

$$E\left(Y_{t+1}|Y_t\right) = \alpha Y_t + \mu_\epsilon$$

Usaremos esta expressão para realizar a previsão de Y_{t+1} dado que conhecemos Y_t. Deve-se notar que esta expressão não é necessariamente u número inteiro. Logo, devemos transformar esta esperança para ser um inteiro não negativo, nesse caso o valor mais próximo do número inteiro, e será denotado por $\hat{Y}_{t+1}$.

2.5 Medidas de Qualidade do Ajuste

O estudo será dividido em duas partes, na primeira em que usaremos para estimar os parâmetros e verificador o desempenho dos métodos de estimação usaremos o viés e o Erro Quadrático Médio (EQM). Considere um conjunto de K simulações de Monte Carlo de amostras de tamanho n do processo INAR(1) e seja $\hat{\theta}^{(i)}$ a estimativa de θ na i-ésima repetição, então o EQM simulado e o viés simulado de $\hat{\theta}$ são dados respectivamente por:

$$EQM(\hat{\theta}) = \frac{1}{K}\sum_{i=1}^{K}\left(\hat{\theta^{(i)}} - \theta\right)^2$$

$$vies(\hat{\theta}) = \theta - \hat{\theta}^{*},$$

em que $\hat{\theta}^{*}$ é a estimativa média de θ, ou seja, $\hat{\theta}^{*} = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}^{(i)}$.

Na segunda parte do estudo, onde faremos a previsão utilizando os parâmetros encontrados, usaremos o MAE (Erro Médio Absoluto) e a Raiz do Erro Quadrático Médio (RMSE), para avaliar o desempenho dos preditores. Estas medidas podem ser encontradas já sendo utilizadas em trabalhos, como em Mahmoudi, Rostami e Roozegar (2018) para a previsão 1 passo à frente. Seja $Y_1, Y_2, \cdots, Y_{n+m}$ uma amostra aleatória de tamanho $n+m$. A amostra será dividida em duas partes. A primeira parte será formada pelas primeiras n observações, e será utilizada para realizar a estimação dos parâmetros α e λ no processo Poisson, e nos outros casos para realizar a estimação de α e μ, já que estamos usando a distribuição binomial negativa e geométrica parametrizada pela média. A segunda parte será formada pelas m observações restantes que são usadas para calcular o MAE e o RMSE, que são respectivamente:

$$\text{MAE} = \frac{1}{m-1}\sum_{i=n}^{n+m-1}|Y_{i+1} - \hat{Y}_{i+1}|$$

$$\text{RMSE} = \sqrt{\frac{1}{m-1}\sum_{i=n}^{n+m-1}(Y_{i+1} - \hat{Y}_{i+1})^2}$$

onde $\hat{Y}_{i+1}$ representa a previsão de Y_{i+1} dado que conhecemos Y_i.

3 Metodologia

Nesta seção descrevem-se os procedimentos metodológicos, no qual as simulações de Monte Carlo, bem como as estimações de parâmetros e previsões do processo INAR(1), foram realizadas utilizando a linguagem de programação **R**. O R, Core Team (2025) é amplamente empregado em análises estatísticas e modelagem de séries temporais devido à sua flexibilidade e à disponibilidade de diversos pacotes especializados.

Foram geradas séries temporais de contagem com diferentes distribuições de erro (Poisson, Binomial Negativa e Geométrica), seguindo a estrutura do modelo INAR(1). Para cada série simulada, os parâmetros do modelo foram estimados pelos métodos de *Mínimos Quadrados Condicionais (MQC)* e *Máxima Verossimilhança Condicional (MVC)*. Com base nos parâmetros estimados, foram realizadas previsões de um passo à frente, permitindo avaliar o desempenho preditivo de cada método.

Para a estimação via máxima verossimilhança condicional, utilizou-se o pacote `optimx` Nash (2014), que fornece funções de otimização robustas e flexíveis para minimizar ou maximizar funções arbitrárias, incluindo funções de log-verossimilhança complexas. Já o pacote `DBI` (R-SIG-DB) (2023) foi empregado para a conexão com bancos de dados via drivers ODBC, permitindo a extração e manipulação de bases de dados externas diretamente no ambiente R.

Dessa forma, todas as etapas da análise, desde a geração das séries até a estimação dos parâmetros e avaliação das previsões, foram realizadas de maneira automatizada e reprodutível, garantindo a consistência e confiabilidade dos resultados obtidos.

Todos os códigos podem ser encontrados em minha página no meu github.

3.1 Estudo de Simulação de Monte Carlo

Experimentação de Monte Carlo significa o uso de valores aleatórios para a estimação de alguma função de uma distribuição de probabilidade. Um problema que não possui um componente estocástico pode ser colocado como um problema com um componente que pode ser identificado como a esperança de uma função de uma variável aleatória (VA). Isso pode ser feito através da decomposição de uma função densidade de probabilidade. O problema é então resolvido pela estimação do valor esperado por meio do uso de uma amostra aleatória da distribuição da Variável Aleatória.

A simulação de Monte Carlo é uma técnica estatística amplamente utilizada para investigar o comportamento de estimadores ou processos estocásticos por meio da repetição de experimentos aleatórios. O método consiste na geração de amostras aleatórias sucessivas, baseadas em distribuições probabilísticas previamente definidas, permitindo a análise empírica da variabilidade dos resultados. Segundo (KROESE et al., 2014), essa abordagem é particularmente útil quando soluções analíticas são impraticáveis ou inexistentes, sendo aplicada em diversas áreas como estatística, física, engenharia e finanças. No campo da inferência estatística, a simulação de Monte Carlo se destaca por fornecer um arcabouço robusto para a avaliação de propriedades assintóticas e de pequeno porte amostral dos estimadores, como viés, consistência, eficiência e erro quadrático médio, sendo especialmente valiosa em modelos complexos ou não lineares (RUBINSTEIN; KROESE, 2016).

Neste trabalho, utilizou-se a simulação de Monte Carlo com o objetivo de estudar o comportamento dos estimadores do modelo INAR(1) sob as três distribuições de inovação. Foram geradas milhares de réplicas de séries temporais sintéticas, com variações sistemáticas nos parâmetros de interesse, permitindo avaliar empiricamente a qualidade das estimativas e a acurácia das previsões obtidas pelos métodos de estimação considerados.

3.1.1 Simulação de Monte Carlo para comparar o desempenho dos estimadores

A simulação tem como objetivo comparar o desempenho dos dois estimadores, Mínimos Quadrados Condicionais (MQC) e Máxima Verossimilhança Condicional (MVC) dos parâmetros α e λ do processo INAR(1), utilizando as três distribuições que vamos comparar. Para isso, foram geradas 5000 amostras do processo INAR(1) para cada distribuição definido na equação (2.1). Para cada amostra gerada calculamos as estimativas de $\widehat{\alpha}_{MQC}$, $\widehat{\alpha}_{MVC}$, $\widehat{\lambda}_{MQC}$ e $\widehat{\lambda}_{MVC}$ com tamanhos amostrais n de 50, 100, 300 e 500. e os seguintes valores dos parâmetros $\alpha = 0.1$ e $\lambda = 5$, $\alpha = 0.3$ e $\lambda = 8$. Para cada parâmetro, $\widehat{\alpha}$ e $\widehat{\lambda}$ de α e λ, respectivamente, calculamos a estimativa média, o viés e o Erro Quadrático Médio, no caso da distribuição *Binomial Negativa*, utilizamos o parâmetro de variabilidade σ^2.

Ou seja, para cada cenário, geraram-se séries temporais com tamanho fixo n e repetição $R = 5000$ de simulações.

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \quad t = 1, 2, \cdots, n$$

A escolha dos valores de parâmetros para a simulação de Monte Carlo baseia-se na literatura consolidada sobre processos INAR(1). Para o coeficiente autorregressivo, foram adotados os valores $\alpha \in \{0.3, 0.5, 0.7\}$, que representam dependência temporal fraca, moderada e forte, respectivamente (JUNG; TREMAYNE, 2003; FREELAND; MCCABE, 2004). Esta configuração é amplamente utilizada em estudos de simulação por proporcionar cobertura adequada do espaço paramétrico relevante, evitando valores extremos que podem causar instabilidade numérica (BU; MCCABE; HADRI, 2008).

Para os parâmetros das distribuições de inovação, foram selecionados $\mu \in \{1.5, 2.5\}$

para a média, valores que se situam no intervalo recomendado por (WEISS, 2008) como representativo da maioria das aplicações empíricas em dados de contagem. A parametrização pela média facilita a comparação direta entre as distribuições com a distribuição de Poisson $\mu = \lambda$.

Ainda foram utilizados duas outras combinações de parâmetros, para observar a influência fraca ou forte de α, ou seja, o quanto a informação anterior afeta de maneira direta o valor da próxima observação, logo, foi acrescentado cenário em que $\alpha = 0,1$ e $\lambda = 5$ e o cenário em que $\alpha = 0,9$ e $\lambda = 4$.

Os tamanhos de amostra quando $n = 500$ proporciona boa aproximação assintótica. O número de 5.000 replicações de Monte Carlo segue a recomendação de (MACKINNON, 2007) para estudos comparativos, garantindo erro padrão suficientemente baixo para conclusões estatísticas robustas com custo computacional razoável.

4 Resultados

4.1 Simulações de Monte Carlo

Nos cenários simulados, observa-se que, de forma geral, os estimadores apresentam melhor desempenho à medida que o tamanho amostral aumenta. Para todos os modelos considerados (Poisson, Binomial Negativa e Geométrica), o viés tende a se aproximar de zero conforme n cresce, indicando consistência dos estimadores. Essa mesma tendência pode ser comprovada pela redução do Erro Quadrático Médio (EQM), que se torna cada vez menor, evidenciando maior precisão nas estimativas a medida que o tamanho de amostra aumenta.

Tabela 1 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,3$ e $\lambda = 1,5$

Amostra	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
Estimativa dos Parâmetros												
$n = 50$	0,2611	1,5777	0,2789	1,5396	0,2609	1,5760	0,2780	1,5488	0,2617	1,5826	0,2987	1,5029
$n = 100$	0,2780	1,5458	0,2867	1,5272	0,2827	1,5339	0,2981	1,5243	0,2802	1,5355	0,3000	1,4936
$n = 300$	0,2926	1,5148	0,2957	1,5082	0,2926	1,5135	0,2932	1,5126	0,2926	1,5162	0,2993	1,5018
$n = 500$	0,2970	1,5055	0,2989	1,5014	0,2956	1,5074	0,2964	1,5057	0,2967	1,5081	0,3005	1,4998
Viés												
$n = 50$	-0,0389	0,0777	-0,0211	0,0396	-0,0391	0,0760	-0,0220	0,0488	-0,0383	0,0826	-0,0013	0,0029
$n = 100$	-0,0220	0,0458	-0,0133	0,0272	-0,0173	0,0339	-0,0109	0,0243	-0,0198	0,0355	0,0000	-0,0064
$n = 300$	-0,0074	0,0148	-0,0043	0,0082	-0,0074	0,0135	-0,0068	0,0126	-0,0074	0,0162	-0,0007	0,0018
$n = 500$	-0,0030	0,0055	-0,0011	0,0014	-0,0044	0,0074	-0,0036	0,0057	-0,0033	0,0081	0,0005	-0,0002
Erro Quadrático Médio												
$n = 50$	0,0202	0,1178	0,0193	0,1115	0,0202	0,1335	0,0168	0,1169	0,0195	0,1679	0,0079	0,1020
$n = 100$	0,0105	0,0600	0,0091	0,0532	0,0101	0,0625	0,0084	0,0553	0,0104	0,0815	0,0040	0,0510
$n = 300$	0,0034	0,0193	0,0028	0,0165	0,0033	0,0201	0,0027	0,0176	0,0035	0,0275	0,0013	0,0169
$n = 500$	0,0020	0,0115	0,0017	0,0098	0,0021	0,0126	0,0017	0,0110	0,0021	0,0164	0,0008	0,0097

Fonte: Elaboração própria

No primeiro cenário, os quais os resultados estão na tabela 1, em que $\alpha = 0,3$ e $\lambda = 1,5$ as estimativas obtidas por Mínimos Quadrados Condicionais (MQC) e por Máxima Verossimilhança Condicional (MVC) se mostraram próximas do valor verdadeiro, especialmente para $n \geq 300$, mostrando de maneira direta que os estimadores $\widehat{\alpha}_{MQC}$, $\widehat{\lambda}_{MQC}$, $\widehat{\alpha}_{MVC}$ e $\widehat{\lambda}_{MVC}$ são consistentes[1]. O viés foi ligeiramente maior para amostras pequenas, em que ($n = 50$), mas reduziu de forma consistente com o aumento da amostra. Entre os modelos, a distribuição geométrica apresentou os menores valores de Erro Quadrático Médio (EQM), sugerindo maior eficiência nesse caso específico.

[1] Um estimador é consistente quando, à medida que o tamanho da amostra aumenta, ele converge em probabilidade para o valor verdadeiro do parâmetro que está sendo estimado. Em outras palavras, $\hat{\theta}_n \xrightarrow{P} \theta$ quando $n \to \infty$.

Tabela 2 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,5$ e $\lambda = 2$

Amostra	Estimativa dos Parâmetros											
	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
$n = 50$	0,4491	2,1913	0,4833	2,0558	0,4473	2,2061	0,4762	2,0977	0,4494	2,1879	0,4983	1,9948
$n = 100$	0,4750	2,0953	0,4923	2,0270	0,4745	2,0914	0,4865	2,0471	0,4744	2,1054	0,5008	2,0007
$n = 300$	0,4899	2,0352	0,4971	2,0068	0,4911	2,0352	0,4933	2,0267	0,4890	2,0424	0,4990	2,0028
$n = 500$	0,4946	2,0213	0,4985	2,0059	0,4950	2,0195	0,4966	2,0135	0,4944	2,0201	0,4997	1,9991
	Viés											
$n = 50$	-0,0509	0,1913	-0,0167	0,0558	-0,0527	0,2061	-0,0238	0,0977	-0,0506	0,1879	-0,0017	-0,0052
$n = 100$	-0,0250	0,0953	-0,0077	0,0270	-0,0255	0,0914	-0,0135	0,0471	-0,0256	0,1054	0,0008	0,0007
$n = 300$	-0,0101	0,0352	-0,0029	0,0068	-0,0089	0,0352	-0,0067	0,0267	-0,0110	0,0424	-0,0010	0,0028
$n = 500$	-0,0054	0,0213	-0,0015	0,0059	-0,0050	0,0195	-0,0034	0,0135	-0,0056	0,0201	-0,0003	-0,0009
	Erro Quadrático Médio											
$n = 50$	0,0189	0,3246	0,0119	0,2044	0,0197	0,3768	0,0122	0,2527	0,0192	0,4300	0,0045	0,1710
$n = 100$	0,0089	0,1493	0,0055	0,0938	0,0088	0,1624	0,0060	0,1180	0,0089	0,1983	0,0022	0,0839
$n = 300$	0,0029	0,0482	0,0017	0,0301	0,0029	0,0553	0,0021	0,0432	0,0029	0,0639	0,0007	0,0285
$n = 500$	0,0017	0,0300	0,0010	0,0183	0,0017	0,0312	0,0012	0,0240	0,0017	0,0373	0,0004	0,0166

Fonte: Elaboração própria

No segundo cenário, no qual temos os parâmetros $\alpha = 0,5$ e $\lambda = 2$, e os resultados expressos na tabela 2, o padrão observado se manteve: maior dispersão para amostras pequenas e convergência para os valores verdadeiros a medida que o tamanho da amostra aumentava. O viés para o parâmetro α foi negativo na maioria dos casos, mas de uma pequena magnitude, enquanto o parâmetro λ apresentou tendência de superestimação para amostras menores. Destaca-se que, para esse cenário, os estimadores de Máxima Verossimilhança Condicional novamente se mostraram ligeiramente superiores em termos de EQM, sobretudo na distribuição geométrica mais uma vez.

Tabela 3 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,7$ e $\lambda = 2,5$

Amostra	Estimativa dos Parâmetros											
	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
$n = 50$	0,6408	2,9617	0,6901	2,5688	0,6376	2,9772	0,6798	2,6551	0,6417	2,9543	0,6994	2,4925
$n = 100$	0,6699	2,7388	0,6953	2,5319	0,6672	2,7638	0,6874	2,6062	0,6676	2,7545	0,0,6991	2,4981
$n = 300$	0,6899	2,5789	0,6981	2,5117	0,6901	2,5805	0,6946	2,5435	0,6904	2,5784	0,6997	2,5017
$n = 500$	0,6934	2,5528	0,6988	2,5080	0,6931	2,5557	0,6949	2,5399	0,6936	2,5550	0,6998	2,5041
	Viés											
$n = 50$	-0,0592	0,4617	-0,0099	0,0688	-0,0624	0,4772	-0,0202	0,1551	-0,0583	0,4543	-0,0006	-0,0075
$n = 100$	-0,0301	0,2388	-0,0047	0,0319	-0,0328	0,2638	-0,00126	0,1062	-0,0324	0,2545	-0,0009	-0,0019
$n = 300$	-0,0101	0,0789	-0,0019	0,0117	-0,0099	0,0805	-0,0054	0,0435	-0,0096	0,0784	-0,0003	0,0017
$n = 500$	-0,0066	0,0528	-0,0012	0,0080	-0,0069	0,0557	-0,0051	0,0399	-0,0064	0,0550	-0,0002	0,0041
	Erro Quadrático Médio											
$n = 50$	0,0159	1,2336	0,0018	0,2511	0,0158	1,0589	0,0066	0,4843	0,0200	0,3592	0,0126	0,2439
$n = 100$	0,0063	0,4322	0,0020	0,1431	0,0068	0,4821	0,0030	0,2366	0,0067	0,5354	0,0008	0,1256
$n = 300$	0,0019	0,01319	0,0007	0,0487	0,0019	0,1386	0,0012	0,0884	0,0020	0,1568	0,0003	0,0419
$n = 500$	0,0011	0,0773	0,0004	0,0279	0,0011	0,0816	0,0012	0,0790	0,0011	0,0893	0,0002	0,0244

Fonte: Elaboração própria

O terceiro cenário, com $\alpha = 0,7$ e $\lambda = 2,5$, os resultados podem ser encontrados na tabela 3, os resultados mais uma vez reforçam a hipótese de melhora da estimativa a medida que o tamanho da amostra aumenta. O parâmetro α foi subestimado nas amostras em que $n \leq 100$, mas essa diferença desaparece progressivamente em $n = 300$ e $n = 500$. Já para λ, houve uma tendência inicial de superestimação, com valores se aproximando do valor verdadeiro à medida em que o tamanho da amostra crescia. Mais uma vez, as estimativas via Máxima Verossimilhança Condicional apresentou valores de EQM menor do que as estimativas usando Mínimos Quadrados Condicionais.

Tabela 4 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,1$ e $\lambda = 5$

Amostra	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
Estimativa dos Parâmetros												
$n = 50$	0,1004	4,9950	0,1077	4,9545	0,1017	4,9899	0,1045	4,9819	0,0972	5,0075	0,1091	4,9407
$n = 100$	0,0964	5,0148	0,1002	4,9940	0,0967	5,0172	0,0980	5,0111	0,0965	5,0168	0,1041	4,9754
$n = 300$	0,0963	5,0213	0,0973	5,0156	0,0973	5,0137	0,0977	5,0120	0,0955	5,0283	0,1017	4,9942
$n = 500$	0,0979	5,0107	0,0985	5,0074	0,0979	5,0105	0,0977	5,0115	0,0970	5,0195	0,1009	4,9981
Viés												
$n = 50$	0,0004	-0,0050	0,0077	-0,0455	0,0017	-0,0101	0,0045	-0,0181	-0,0028	0,0075	0,0091	-0,0593
$n = 100$	-0,0036	0,0148	0,0002	-0,0060	-0,0033	0,0172	-0,0020	0,0111	-0,0035	0,0168	0,0041	-0,0246
$n = 300$	-0,0037	0,0213	-0,0027	0,0156	-0,0027	0,0137	-0,0023	0,0120	-0,0045	0,0283	0,0017	-0,0058
$n = 500$	-0,0021	0,0107	-0,0015	0,0074	-0,0021	0,0105	-0,0023	0,0115	-0,0030	0,0195	0,0009	-0,0019
Erro Quadrático Médio												
$n = 50$	0,0110	0,4470	0,0130	0,5071	0,0111	0,4807	0,0114	0,4936	0,0111	0,9609	0,0036	0,6929
$n = 100$	0,0071	0,2656	0,0076	0,2830	0,0071	0,2885	0,0071	0,2870	0,0071	0,5320	0,0018	0,3683
$n = 300$	0,0031	0,1120	0,0031	0,1130	0,0031	0,1186	0,0030	0,1147	0,0031	0,1993	0,0006	0,1190
$n = 500$	0,0020	0,0698	0,0020	0,0701	0,0020	0,0758	0,0019	0,0718	0,0020	0,1227	0,0003	0,0714

Fonte: Elaboração própria

No quarto cenário, em que $\alpha = 0,1$ e $\lambda = 5$, os resultados encontram-se na tabela 4, verifica-se uma particular estabilidade mesmo em tamanhos amostrais reduzidos. As estimativas de α ficaram bastante próximas do valor verdadeiro, e o viés foi praticamente nulo já a partir de $n = 100$. Para o parâmetro λ, o viés foi ligeiramente positivo, mas de magnitude pequena. O EQM também se mostrou bastante reduzido nesse cenário, sugerindo maior facilidade de estimação quando os parâmetros assumem esses valores.

Por fim, no cenário mais extremo, com $\alpha = 0,9$ e $\lambda = 4$, os resultados estão descritos na tabela 5, e conforme observado indicam maior dificuldade em obter boas estimativas em amostras pequenas. O viés em λ foi substancial para $n = 50$, mas reduziu gradualmente em amostras maiores. O Erro Quadrático Médio também foi elevado nesse caso, sobretudo na distribuição geométrica com $n = 50$, refletindo maior instabilidade. Contudo, o resultado de convergência se manteve e no caso em que $n = 500$ as estimativas já se aproximam consideravelmente dos valores verdadeiros e observamos que para α via Máxima Verossimilhança Condicional, o erro quadrático médio é praticamente zero, independente do tamanho da amostra.

Tabela 5 – Estimativas dos parâmetros, viés e Erro Quadrático Médio quando $\alpha = 0,9$ e $\lambda = 4$

Amostra	Poisson				Binomial Negativa				Geométrica			
	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$	$\widehat{\alpha}_{MQC}$	$\widehat{\lambda}_{MQC}$	$\widehat{\alpha}_{MVC}$	$\widehat{\lambda}_{MVC}$
Estimativa dos Parâmetros												
$n = 50$	0,8806	4,6253	0,8973	4,0770	0,8809	4,6211	0,9065	4,3938	0,8546	5,4839	0,9000	3,9802
$n = 100$	0,8857	4,5137	0,8984	4,0517	0,8850	4,5393	0,9032	4,1635	0,8729	4,9787	0,9000	3,9913
$n = 300$	0,8927	4,2786	0,8994	4,0178	0,8921	4,3020	0,9015	4,0217	0,8892	4,4178	0,8998	4,0051
$n = 500$	0,8944	4,2136	0,8996	4,0105	0,8943	4,2225	0,9004	4,0733	0,8930	4,2787	0,9000	4,0029
Viés												
$n = 50$	-0,0194	0,6253	-0,0027	0,0770	-0,0191	0,6211	0,0065	0,3938	-0,0454	1,4839	0,0000	-0,0198
$n = 100$	-0,0143	0,5137	-0,0016	0,0517	-0,0150	0,5393	0,0032	0,1635	-0,0271	0,9787	0,0000	-0,0087
$n = 300$	-0,0073	0,2786	-0,0006	0,0178	-0,0079	0,3020	0,0015	0,0217	-0,0108	0,4178	-0,0002	0,0051
$n = 500$	-0,0056	0,2136	-0,0004	0,0105	-0,0057	0,2225	0,0004	0,0733	-0,0070	0,2787	0,0000	0,0029
Erro Quadrático Médio												
$n = 50$	0,0019	1,9345	0,0005	0,5100	0,0018	1,8906	0,0010	1,8287	0,0066	7,7092	0,0002	0,5542
$n = 100$	0,0011	1,3793	0,0002	0,2789	0,0011	1,4687	0,0006	0,7740	0,0027	3,8210	0,0001	0,2771
$n = 300$	0,0005	0,7112	0,0001	0,1127	0,0005	0,7348	0,0002	0,2791	0,0008	1,2659	0,0000	0,0968
$n = 500$	0,0003	0,4846	0,0000	0,0688	0,0003	0,5139	0,0002	0,2555	0,0004	0,7191	0,0000	0,0573

Fonte: Elaboração própria

De modo geral, os resultados das simulações apontam que tanto Mínimos Quadrados Condicionais, quanto a Máxima Verossimilhança Condicional são consistentes, mas a MVC tende a apresentar menor EQM, principalmente observando a distribuição Geométrica. Além disso o aumento do tamanho amostral reduz de forma significativa tanto o viés quanto a variabilidade das estimativas, confirmando as propriedades assintóicas dos estimadores.

4.2 Resultados da Aplicação com dados reais

5 Considerações finais

O presente trabalho teve como finalidade avaliar modelos de aprendizado de máquina na tentativa de prever o próximo dia de compra de um cliente. Pode-se dizer que, com os resultados obtidos dos modelos avaliados, as técnicas mostraram-se eficazes para prever o próximo dia de compra dos clientes, apresentando resultados satisfatórios para o banco de dados utilizado. Apesar dos desafios iniciais, relacionados ao sobreajuste do modelo aos dados de treinamento, o uso da engenharia de características permitiu alcançar um resultado satisfatório, deixando de aprendizado a real importância dos passos iniciais no treinamento de um modelo. Vale deixar claro que, apesar do modelo *Multilayer Perceptron* ter apresentado o melhor resultado, isso não implica dizer que ele é de fato o melhor modelo, e sim, de que ele foi bom em se ajustar à esses dados em questão. Além disso, exceto pelo modelo SVM, todos os demais modelos apresentaram resultados satisfatórios.

Um ponto importante nesse trabalho que vale se destacar foi a junção da segmentação de clientes RFV com o método de aprendizagem não supervisionada, *K-means*, na qual retorna uma informação importante para o setor comercial. Com base nessa segmentação, as empresas podem ter um controle maior acerca de seus clientes, a possibilidade de personalizar suas estratégias de marketing e comunicação para atender às necessidades específicas de cada grupo de clientes, garantem uma maximização do retorno sobre investimento. A respeito do real objetivo do trabalho, previsão do próximo dia de compra, essa informação por si só possibilitaria o desenvolvimento de ofertas, promoções e mensagens especiais programadas para incentivar a repetição de compras e aumentar o valor do cliente. Além disso, esse dado possibilita um gerenciamento mais eficiente do estoque, evitando excesso ou falta de estoque.

Em suma, os resultados do estudo destacam o potencial da utilização do aprendizado de máquina no setor comercial. Além da previsão do próximo dia de compra, as técnicas podem ser adaptadas para identificar outros comportamentos, como preferências de produtos e ciclos de vida do cliente. Em geral, é importante reconhecer as limitações deste estudo, reiterando que um bom modelo requer boas informações, além-claro, das melhorias no pré-processamento: seleção de variáveis e tratamento das informações.

Por fim, em um certo ponto da metodologia, foi mencionado que algumas observações apresentaram variável resposta faltante, decorrente do fato de alguns clientes não voltarem a realizar compras. Esse tipo de comportamento é comum no campo da análise e sobrevivência. A área lida especificamente com os chamados “dados censurados”, em que a variável resposta, nesse caso uma compra subsequente, pode ser faltante para algumas observações. Futuras pesquisas podem abordar essas limitações para aprimorar esses modelos, com técnicas de análise e sobrevivência ou até mesmo remodelar para um problema de classificação, buscando identificar clientes que deixarão de comprar em uma janela de tempo pré-definida.

Referências

AL-OSH, M. A.; ALZAID, A. A. First-order integer-valued autoregressive (inar(1)) process. *Journal of Time Series Analysis*, v. 8, n. 3, p. 261–275, 1987.

BARCELOS, B. I. *Teste da raiz unitária Dickey-Fuller no modelo INAR(1)*. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2008.

BU, R.; MCCABE, B.; HADRI, K. Maximum likelihood estimation of higher-order integer-valued autoregressive processes. *Journal of Time Series Analysis*, Wiley Online Library, v. 29, n. 6, p. 973–994, 2008.

CAMERON, A. C.; TRIVEDI, P. K. *Regression Analysis of Count Data*. [S.l.]: Cambridge University Press, 1998.

DU, J. G.; LI, Y. The integer-valued autoregressive (inar(p)) model. *Journal of Time Series Analysis*, v. 12, n. 2, p. 129–142, 1991.

FOKIANOS, K.; TJøSTHEIM, D. Poisson autoregression. *Journal of the American Statistical Association*, v. 104, n. 488, p. 1430–1439, 2009.

FREELAND, R. K. *Statiscial analysis of discrete time series with application to the analysis of workers compensation claims data*. Tese (Doutorado) — University of British Columbia, 1998.

FREELAND, R. K.; MCCABE, B. P. M. Analysis of low count time series data by Poisson autoregression. *Journal of Time Series Analysis*, Wiley Online Library, v. 25, n. 5, p. 701–722, 2004.

GOMES, K. S. *Modelagem INAR(p) para previsão de índices de qualidade do ar*. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, 2009.

JUNG, R. C.; TREMAYNE, A. R. Testing for serial dependence in time series models of counts. *Journal of Time Series Analysis*, Wiley Online Library, v. 24, n. 1, p. 65–84, 2003.

KROESE, D. P. et al. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley, v. 6, n. 6, p. 386–392, 2014.

LATOUR, A. Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis*, v. 19, n. 4, p. 439–455, 1998.

MACKINNON, J. G. Bootstrap hypothesis testing. In: *Handbook of Statistics*. [S.l.]: Elsevier, 2007. v. 26, p. 183–213.

MAHMOUDI, E.; ROSTAMI, M.; ROOZEGAR, R. A new integer-valued ar(1) process based on power series thinning operator. *Communications in Statistics - Simulation and Computation*, v. 47, n. 10, p. 2895–2906, 2018.

MCKENZIE, E. Some simple models for discrete variate time series. *Water Resources Bulletin*, v. 21, n. 4, p. 645–650, 1985.

MCKENZIE, E. Autoregressive moving-average processes with negative binomial and geometric marginal distributions. *Advances in Applied Probability*, v. 18, n. 3, p. 679–705, 1986.

MCKENZIE, E. Discrete variate time series. In: FINKENSTäDT, B.; ROOTZéN, H. (Ed.). *Seminar on Stochastic Analysis, Random Fields and Applications III*. [S.l.]: Springer, 2003. p. 305–312.

NASH, J. C. *optimx: Extended Replacement and Comparison of Optimization Methods*. [S.l.], 2014. Disponível em: <https://CRAN.R-project.org/package=optimx>.

PEDELI, X.; KARLIS, D. A bivariate inar(1) model for count time series. *Statistical Modelling*, v. 11, n. 1, p. 35–53, 2011.

R, Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2025. Disponível em: <https://www.R-project.org/>.

(R-SIG-DB), R. S. I. G. on D. *DBI: R Database Interface*. [S.l.], 2023. Disponível em: <https://CRAN.R-project.org/package=DBI>.

RUBINSTEIN, R. Y.; KROESE, D. P. *Simulation and the Monte Carlo Method*. 3. ed. [S.l.]: John Wiley & Sons, 2016.

SILVA, I. M. M. d. *Contributions to the analisys of discrete-valued time series*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2005.

STEUTEL, F. W.; HARN, K. van. Discrete analogues of self-decomposability and stability. *The Annals of Probability*, v. 7, n. 5, p. 893–899, 1979.

WEISS, C. H. Thinning operations for modeling time series of counts—a survey. *AStA Advances in Statistical Analysis*, Springer, v. 92, n. 3, p. 319–341, 2008.

ZHANG, J.; WANG, M.; ZHU, F. Modeling monthly tuberculosis incidence with inar models. *Statistics in Medicine*, v. 38, n. 11, p. 1953–1966, 2019.