REVISTA BRASILEIRA

DE ESTATÍSTICA

volume 80

número 247

Janeiro/dezembro 2022

IBGE
Instituto Brasileiro de Geografia e Estatística

Instituto Brasileiro de Geografia e Estatística – IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

Volume 80 - número 247 – jan/dez 2022

ISSN 2675-3243

R. Bras. Estat., Rio de Janeiro, v.80, n.247, p. 1-91, jan/dez 2022

Instituto Brasileiro de Geografia e Estatística – IBGE

Revista Brasileira de Estatística, ISSN 2675-3243

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site https://rbesojs.ibge.gov.br.

Os artigos submetidos à RBE não devem ter sido publicados em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração

José André de Moura Brito

Capa

José André de Moura Brito

Informações Adicionais

Revista Brasileira de Estatística/IBGE - v.1, n.1 (jan/mar. 1940) – Rio de Janeiro: IBGE, 1940.v. trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística. Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989). Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN publicação online 2675-3243, a partir de 2019.

I. Estatística – Periódicos. II. IBGE. III. Associação Brasileira de Estatística.

Gerência de Biblioteca e Acervos Especiais CDU 31(05) RJ-IBGE/88-05 (ver. 2009) PERIÓDICO

Sumário

Nota do Editor

Sejam bem-vindos a mais um número da Revista Brasileira de Estatística.

Esta edição marca uma importante mudança da revista para o sistema OJS hospedado agora no IBGE. Agradecemos a todos envolvidos no processo de mudança liderado pelo nosso colega Editor de revista, José André Brito. Estamos otimistas que a mudança trará mais estabilidade e dinamismo aos processos de submissão e avaliação.

Quatro trabalhos compõem este número da revista.
No primeiro, Isabela Lima e Carla Brighenti trata do desenvolvimento de testes de hipóteses convencional e sequencial sobre o parâmetro da exponencial.

O comportamento temporal do índice de expansão do comércio (IEC) na cidade de São Paulo no período de junho de 2011 a julho de 2020 é avaliado no trabalho de Luciane Giarola usando modelos e métodos de séries temporais.

Vicente Marques contrasta resultados dos censos agropecuários de 2006 e 20017 discutindo condicionantes, por exemplo, climáticos, com atenção à agricultura familiar.

O trabalho de Gabriel Alves e coautores destaca bons resultados preditivos ao adotar métodos de regressão quantílica na análise da letalidade da covid-19 nos municípios do Estado de São Paulo.

Agradecemos as contribuições dos autores e a revisão efetuada por editores de seção e revisores.

Desejamos uma boa leitura!

Paulo Justiniano Ribeiro Jr.

Editor RBE.

Revista Brasileira de Estatística, V.80, N. 247, p.7-27, 2022

COMPARAÇÃO ENTRE OS TESTES DE HIPÓTESES CONVENCIONAL E SEQUENCIAL SOBRE O PARÂMETRO DA EXPONENCIAL

Isabela da Silva Lima

isabela_lima30@hotmail.com

Universidade Federal de Lavras

Carla Regina Guimarães Brighenti

carlabrighenti@ufsj.edu.br

Universidade Federal de São João del-Rei / Universidade Federal de Lavras

Resumo:
Os testes de hipóteses são regras de decisão sobre os parâmetros de uma distribuição baseadas em informações contidas na amostra. Os problemas mais usuais de decisões sobre estes parâmetros consideram a distribuição Normal. Neste trabalho, são apresentados os conceitos que estruturam um Teste de Hipóteses e, de modo particular, foi estudada a teoria para o Teste de Hipóteses Convencional sobre o parâmetro da distribuição Exponencial. Além disso, foi estudado o Teste Sequencial da Razão de Probabilidade, no qual o tamanho amostral é considerado uma variável aleatória, também considerando a distribuição Exponencial. Elaborou-se um script para a metodologia do teste de hipóteses Convencional e Sequencial utilizando o software R, sendo a teoria exemplificada para um problema de controle de qualidade, em que os dados foram simulados. O Teste de Hipóteses Sequencial resultou na mesma decisão que o Convencional, entretanto, com um tamanho de amostra menor, mostrando-se mais vantajoso.
Palavras-chave: Amostragem Sequencial; distribuição Gama, SPRT

Abstract:
Hypothesis Tests are decision rules about the parameters of a distribution based on information contained in the sample. The most common decision problems about these parameters are the Normal distribution. This paper, are presented the concepts that structure a Hypothesis Test, in particular, the theory for the Conventional Hypothesis Test on the Exponential distribution parameter was studied. Besides that, the Sequential Probability Ratio Test was studied, in which the sample size is considered a random variable. A study of general theory was carried out and then an approach to Exponential distribution. A script for the Conventional and Sequential hypothesis testing methodologies was developed using the R software. The theory studied was exemplified for a quality control problem, in which the data were simulated. The Sequential Hypothesis Test resulted in the same decision as the Conventional, however, with a smaller sample size, exhibiting to be more advantageous.

Keywords: Sequential Sampling; Gama Distribution; SPRT

1. INTRODUÇÃO

Para a realização de teste de hipóteses especificam-se duas hipóteses, denominadas hipótese nula (H_0) e hipótese alternativa (H_1), e um critério para a rejeição da hipótese nula. Assim, o teste fornece um procedimento para tomar uma decisão sobre rejeitar ou não a hipótese nula sobre o valor dos parâmetros populacionais, por meio dos resultados de uma amostra, geralmente com tamanho fixado antes do experimento.

No entanto, Wald (1947) propôs uma metodologia alternativa para o teste de hipóteses convencional, denominada teste de hipóteses sequencial, que se caracteriza por envolver o tamanho da amostra uma variável determinada pelos dados observados, ou seja, o tamanho da amostra corresponde a uma variável aleatória. Assim, em um teste sequencial, determina-se, a partir da razão de probabilidade após cada observação realizada, se as informações disponíveis são suficientes para rejeitar ou não a hipótese nula.

É importante ressaltar que, vários autores ao considerarem uma abordagem sequencial para um teste, ao contrário da abordagem convencional, conseguiram uma redução no tamanho da amostra com consequente diminuição nos custos de avaliação (Penteado et al. 2008; Liu et al., 2009; Fernandes et al., 2002.)

Percebe-se também que, tanto os autores citados, quanto a maioria dos livros de estatística básica, apresentam os Testes de Hipóteses, tanto o convencional, assim como o sequencial apenas para os parâmetros das distribuições Normal e Binomial. No entanto, ao descrever tempo de vida, em áreas como análise de confiabilidade e sobrevivência uma importante distribuição considerada é a Exponencial. Por exemplo, Leal e Andrade (2018) ao modelar o tempo de vida em horas de um caminhão fora de estrada utilizado na mineração, por meio da análise de confiabilidade paramétrica, encontraram o melhor ajuste utilizando a distribuição Exponencial. Poderia ser interessante testar o tempo de vida de duas frotas de empresas diferentes. Da mesma forma, a distribuição Exponencial também é utilizada para modelar comportamentos de falha em sistemas eletrônicos, tais como duração de lâmpadas, como pode ser visto em Macedo (2018), Caetano e Louzada-Neto (2007).

Desse modo, objetivou-se descrever formalmente os conceitos envolvidos no Teste de Hipóteses Convencional e Sequencial sobre o parâmetro da distribuição Exponencial, e aplicá-los em um exemplo de controle

de qualidade para comparar qual das duas metodologias abordadas é mais vantajosa para o caso apresentado.

2.REFERENCIAL TEÓRICO

2.1 TESTE DE HIPÓTESES

Um dos problemas básicos da Inferência Estatística é o de verificar afirmações sobre características de uma população feitas a partir de dados amostrais. Assim, o Teste de Hipóteses proporciona um método para averiguar se os dados de uma amostra contrariam ou não uma afirmação feita sobre a população (Bussab & Morettin, 2017).

Essas afirmações são chamadas de hipóteses estatísticas. Seja Θ um espaço paramétrico, uma hipótese estatística H é simplesmente uma especificação de um subconjunto de Θ. Se uma hipótese estatística determina totalmente a distribuição, isto é, especifica um subconjunto com apenas um ponto de Θ ela é chamada de hipótese simples e caso contrário, é chamada de hipótese composta (Mood et al., 1974).

Seja θ um parâmetro populacional e θ_0 o verdadeiro valor de θ. Então, uma hipótese simples é da forma: $H:\theta=\theta_0$ e composta unilateral: $H:\theta>\theta_0$ ou $H:\theta<\theta_0$, e ainda tem-se a hipótese composta bilateral que assume a forma: $H:\theta\neq\theta_0$.

Para a realização do teste de hipóteses, explicitam-se duas hipóteses, denominadas de hipótese nula, denotada por $H_0:\{\theta\in\Theta_0\}$, esta é uma hipótese simples, que sugere um valor para o parâmetro populacional ou ainda, representa uma igualdade do parâmetro a ser testado, ou seja, $H_0:\theta=\theta_0$. E hipótese alternativa, denotada por $H_1:\{\theta\in\overline{\Theta}_0\}$, com $\Theta_0\cup\overline{\Theta}_0=\Theta$ Esta pode ser composta: $H_1:\theta>\theta_0$ ou $H_1:\theta<\theta_0$, precisa: $H_1:\theta\neq\theta_0$, ou ainda simples: $H_1:\theta=\theta_1$. Neste caso, as duas hipóteses não serão complementares, já que $H_1:\theta=\theta_1$, onde $\theta\in\Theta_1$ (Mood et al., 1974).

Depois de observar uma determinada amostra, o pesquisador deve decidir se rejeita ou não H_0. Para isto, utiliza-se o teste de hipóteses, que segundo Mood et al. (1974), pode ser definido como:

Seja Υ um teste de uma hipótese estatística, rejeita-se H_0 se e somente se a amostra $(x_1,...,x_n)\in C_\Upsilon$, em que C_Υ é um subconjunto do espaço amostral das observações; C_Υ é chamado de região crítica ou de rejeição do teste Υ.

Desse modo, um teste de hipótese é uma decomposição do espaço amostral das realizações $\mathcal{X}^n$ em C_Υ e $\overline{C}_\Upsilon$ (complementar de C_Υ), de modo que se $(x_1,...,x_n) \in C_\Upsilon$, H é rejeitada, em que $\mathcal{X}^n$ é o espaço amostral. Assim, um teste de hipóteses fica determinado quando especificamos sua região crítica (Mood et al., 1974).

Desta forma, uma decisão será tomada a partir da construção da região crítica. Se o valor observado pertencer a essa região, então rejeita-se H_0, caso contrário não se rejeita H_0. A região crítica é sempre construída sob a hipótese de H_0 ser verdadeira (Bussab & Morettin, 2017).

Ao tomar uma decisão de rejeitar ou não uma hipótese nula, pode-se cometer dois tipos de erros: erro tipo I, rejeita-se a hipótese nula H_0, quando de fato ela é verdadeira e erro tipo II, não rejeita-se H_0 quando de fato ela é falsa. O tamanho do erro tipo I (α) é definido como sendo a probabilidade de se cometer esse erro; do mesmo modo, o tamanho do erro tipo II (β) é a probabilidade de se cometer o erro tipo II (Bussab & Morettin, 2017).

O poder do teste é a probabilidade de rejeitar H_0 quando H_0 é realmente falsa, ou seja, o poder de um teste é igual a $1-\beta$. A função poder do teste Υ associa a cada valor de θ a probabilidade $\pi_\Upsilon(\theta)$ de rejeitar H_0, ela é definida como $\pi_\Upsilon(\theta) = P_\theta[\text{rejeitar } \mathrm{H}_0] = P_\theta[(\mathrm{X}_1,...,\mathrm{X}_n) \in \mathrm{C}_\Upsilon]$, em que C_Υ é a região crítica do teste Υ (Bussab & Morettin, 2017).

O ideal é que ambas as probabilidades dos erros sejam pequenas, mas geralmente, ao se tentar diminuir a probabilidade do erro tipo I, não se consegue controlar o valor da probabilidade do erro tipo II. Para um tamanho de amostra fixo, geralmente é difícil tornar ambos os tipos de probabilidades de erros arbitrariamente pequenos (Casella & Berger, 2002).

Desta forma, fixa-se α em uma probabilidade pequena, usualmente em 0,05 ou 0,01. A justificativa para se fixar o tamanho do erro tipo I é que, normalmente, as hipóteses são elaboradas de forma que a hipótese nula seja mais importante do que a hipótese alternativa; sendo assim, a ocorrência do erro tipo I, recusar a hipótese nula sendo esta verdadeira, é mais grave do que o erro em se recusar a hipótese alternativa sendo esta verdadeira.

De acordo com Mood et al. (1974), o tamanho do teste Υ é definido como o tamanho da região crítica. Um teste Υ^* de $H_0 : X \sim \mathrm{f}_0(.;\theta_0)$ contra $H_1 : X \sim \mathrm{f}_1(.;\theta_1)$, sendo θ_0 e θ_1 conhecidos, é dito ser um teste mais poderoso de tamanho α $(0 < \alpha < 1)$ se e somente se:

(i) $\pi_{\Upsilon^*}(\theta_0) = \alpha$

(ii) $\pi_{\Upsilon^*}(\theta_1) \geq \pi_\Upsilon(\theta_1)$ para qualquer outro teste Υ para o qual $\pi_\Upsilon(\theta_0) \leq \alpha$

Desse modo, um teste Υ^* é mais poderoso de tamanho α se ele possui o tamanho de seu erro tipo I igual a α e tem menor tamanho do erro tipo II sobre todos os outros testes com tamanho de erro tipo I igual a α ou menor.

2.2 TESTE DA RAZÃO DE VEROSSIMILHANÇAS SIMPLES

Seja uma amostra aleatória $X_1,...,X_n$ originada de uma dentre duas distribuições completamente especificadas $f_0(x)$ e $f_1(x)$, o objetivo é decidir de qual delas vem à amostra. Supondo hipóteses simples, o espaço paramétrico Θ contém apenas dois pontos, θ_0 e θ_1. Desse modo, testa-se as hipóteses:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases} \quad (1)$$

Para encontrar procedimentos de testes, ou seja, a estatística do teste, o método da razão de verossimilhanças é bastante utilizado (Casella & Berger, 2002).

A função de verossimilhança é definida como:

$$L(\theta \mid x_1,...,x_n) = L(\theta \mid x) = f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \quad (2)$$

Assim, um teste Υ de $H_0 : \mathrm{X} \sim \mathrm{f}_0(.;\theta_0)$ contra $H_1 : \mathrm{X} \sim \mathrm{f}_1(.;\theta_1)$ é definido como sendo um teste de razão de verossimilhanças simples se Υ é definido por:

- Rejeita-se H_0 se $B \le b_c$
- Não se rejeita H_0 se $B > b_c$

Em que b_c é uma constante não-negativa, chamado de valor crítico e B é a razão de verossimilhanças, dada por:

$$B = B(x_1,...,x_n) = \frac{\prod_{i=1}^{n} f_0(x_i)}{\prod_{i=1}^{n} f_1(x_i)} = \frac{L(\theta_0; x_1,...,x_n)}{L(\theta_1; x_1,...,x_n)} = \frac{L_0}{L_1} \quad (3)$$

Onde L_j é a função de verossimilhança para a amostra de densidade $f_j(\cdot)$, para o caso de amostras independentes. Para cada valor de b_c tem-se um teste diferente.

A importância dos testes de razão de verossimilhanças se dá em consequência do Lema de Neyman-Pearson, citado em Mood et al. (1974):

> *Seja $X_1,\ldots,X_n$ uma amostra aleatória de uma função densidade de probabilidade $f(x;\theta)$, em que θ representa um dos dois valores conhecidos θ_0 ou θ_1, e seja $0<\alpha<1$ fixado. Seja b_c uma constante positiva e C^* um subconjunto de $\mathbb{X}^n$, tal que satisfaça:*

(i) $P_{\theta_0}\left[\left(X_1,...,X_n\right)\in C^*\right]=\alpha$

(ii) $B(x_1,...,x_n)=\dfrac{L\left(\theta_0;x_1,...,x_n\right)}{L\left(\theta_1;x_1,...,x_n\right)}=\dfrac{L_0}{L_1}\le b_c$ *se* $(x_1,...,x_n)\in C^*$ *e* $B(x_1,...,x_n)>\mathrm{b}_c$ *se* $(x_1,...,x_n)\in \overline{C}^*$

Então, o teste Υ^* correspondente à região crítica C^* é o teste mais poderoso de tamanho α relativo ao teste de hipóteses simples dado por $H_0:\theta=\theta_0$ versus $H_1:\theta=\theta_1$.

Para se calcular o tamanho de um teste é necessária uma integração múltipla sobre a região crítica, que pode ser muito difícil.

$$P\left[\left(X_1,...,X_n\right)\in C_\Upsilon\right]=\int...\int_{C_\Upsilon}\prod f(x_i;\theta)dx_i \tag{4}$$

Nos testes de razão de verossimilhanças a região crítica do teste é geralmente dada em termos de uma desigualdade que depende da razão de verossimilhanças, a estatística $B(X_1,....,X_n)$. Um método alternativo a este procedimento é obter a função densidade de probabilidade da estatística $B(X_1,....,X_n)$ sob H_0, $f_B^0(b)$, e fazer a integração simples $\int_0^{b_c} f_B^0(b)db$. Tal procedimento é, em geral, também bastante complexo. Desse modo, o usual é obter uma desigualdade equivalente em termos de uma outra estatística suficiente S, para o qual seja mais fácil obter a distribuição de probabilidade. Assim,

$$B(X)\le b_c \Leftrightarrow S\le k \tag{5}$$

Seja θ um parâmetro de interesse, o princípio da suficiência estabelece que uma estatística $S(X)$ é dita suficiente para o parâmetro θ se ela captura toda a informação sobre θ contida na amostra, e não depende do parâmetro θ (Casella & Berger, 2002).

2.3 TESTE DE HIPÓTESES SEQUENCIAL

O teste sequencial é um teste de hipóteses em que o número de observações que o plano amostral requer não é conhecido antecipadamente, este caracteriza-se por envolver amostras de tamanho variável determinado pelos dados observados. Neste teste, a decisão de rejeitar ou não hipótese é realizada a cada passo, assim, o tamanho total da amostra depende da informação acumulada a cada observação, diferente da amostragem convencional, que possui um número fixo pré-determinado do tamanho da amostra (Barbosa, 1992).

Como a decisão de terminar a amostragem e tomar uma decisão depende dos resultados obtidos em cada passo, o teste sequencial pode diminuir o tempo e os custos da análise em relação aos testes convencionais.

Geralmente, os testes que requerem um tamanho de amostra pré-fixado necessitam de um grande número de observações, o que resulta em custos elevados, além de exigir mais tempo para sua realização. Na amostragem sequencial, as amostras têm em média um terço do tamanho utilizado na amostragem de tamanho fixo, evitando assim desperdícios com medições desnecessárias. Desse modo, o teste sequencial pode ser mais vantajoso (Fonseca, 2008).

O método de teste sequencial foi desenvolvido inicialmente por Wald (1947) e se baseia na razão de probabilidade para determinar, após cada observação feita, se as informações disponíveis são suficientes para rejeitar ou não a hipótese nula (H_0). A função $f_0(x_i,k)$ representa a função probabilidade do resultado amostral, $x_i=(x_1,...,x_k)$ quando H_0 é verdadeira e $f_1(x_i,k)$ representa a função probabilidade quando H_1 é verdadeira.

O teste de razão de probabilidade verifica a seguinte razão, chamada de razão de verossimilhança:

$$R_k=\frac{f_1(x_i,k)}{f_0(x_i,k)} \tag{6}$$

Valores de limites A e B são escolhidos para satisfazer os tamanhos de erros tipo I e II pré-definidos para o teste de hipóteses. Usando α e β para representar a probabilidade desses erros, respectivamente, A e B são calculados de acordo com:

$$\text{A} = \frac{\beta}{(1-\alpha)} \quad \text{e} \quad \text{B} = \frac{(1-\beta)}{\alpha} \tag{7}$$

Assim, o teste sequencial examina as unidades amostrais em sequência até que os resultados são comparados com limites previamente determinados, da seguinte maneira (Wald, 1947):

- Se $\frac{f_1(x_i,k)}{f_0(x_i,k)} \leq \frac{\beta}{(1-\alpha)}$, não se rejeita H_0;
- Se $\frac{f_1(x_i,k)}{f_0(x_i,k)} \geq \frac{(1-\beta)}{\alpha}$, rejeita-se H_0;
- Se $\frac{\beta}{(1-\alpha)} < \frac{f_1(x_i,k)}{f_0(x_i,k)} < \frac{(1-\beta)}{\alpha}$, inconclusivo, continua-se o experimento.

Logo, a partir da primeira observação uma decisão deve ser tomada no sentido de rejeitar ou não a hipótese nula, ou então a de se continuar com a amostragem. Caso o teste for inconclusivo, outra unidade amostral é examinada e, com base nos resultados acumulados dos dois itens da amostra chega-se a uma das três decisões novamente. O processo se prolonga até que

seja tomada a decisão de não rejeitar ou rejeitar a hipótese nula (Brighenti et al., 2013)

2.3 DISTRIBUIÇÃO EXPONENCIAL

A distribuição exponencial é uma distribuição contínua de probabilidade amplamente aplicada em confiabilidade de sistemas, tempos de sobrevivência, tempos de falha, entre outros (Bussab & Morettin, 2017).

Sua função densidade de probabilidade é dada por:

$$f(x;\theta)=\begin{cases}\theta e^{-\theta x}, & \text{se } x\geq 0\\ 0, & \text{se } x<0\end{cases} \tag{8}$$

Em que $\theta>0$ é o parâmetro de taxa da distribuição, e escreve-se $X \sim \text{Exp}(\theta)$, para indicar que X tem distribuição exponencial de parâmetro θ.

O valor esperado e a variância da variável Exponencial X são dados por:

$$E(X)=\frac{1}{\theta} \qquad \text{e} \qquad V(X)=\frac{1}{\theta^2} \tag{9}$$

A distribuição exponencial acumulada é dada por:

$$\text{F}(x;\theta)=\begin{cases}1-e^{-\theta x}, & \text{se } x\geq 0\\ 0, & \text{se } x<0\end{cases} \tag{10}$$

A distribuição exponencial é um caso particular da distribuição Gama, cuja função de densidade de probabilidade é dada por:

$$f(x;n,\theta)=\begin{cases}\dfrac{1}{\Gamma(n)}x^{n-1}\theta^n e^{-\theta x}, & \text{se } x\geq 0\\ 0, & \text{se } x<0\end{cases} \tag{11}$$

Onde $n>0$ é o parâmetro de forma e $\theta>0$, parâmetro de escala, e escreve-se $X \sim \text{Gama}(n,\theta)$, para indicar que X tem distribuição Gama com parâmetros n e θ. Assim, para $n=1$ a função densidade de probabilidade da distribuição Gama se reduz à distribuição Exponencial. A função $\Gamma(n)$ é a função gama, em que $\Gamma(n)=(n-1)!$, sendo n inteiro positivo.

O valor esperado e a variância da variável Gama X são definidos por:

$$E(X)=\frac{n}{\theta} \qquad \text{e} \qquad V(X)=\frac{n}{\theta^2} \tag{12}$$

Em muitos casos é comum encontrar outra parametrização da distribuição exponencial, dada por:

$$f\left(x;\frac{1}{\lambda}\right)=\begin{cases}\dfrac{1}{\lambda}e^{-\frac{x}{\lambda}}, & \text{se } x\geq 0\\ 0, & \text{se } x<0\end{cases} \tag{13}$$

Assim, diz-se que λ é o parâmetro de escala da distribuição e este é o inverso do parâmetro taxa, do mesmo modo denota-se por $X \sim \text{Exp}(\lambda)$. Desse modo, é importante verificar qual das duas especificações está sendo utilizada quando escreve-se $X \sim \text{Exp}(a)$, se a está se referindo ao parâmetro taxa ou ao parâmetro escala da distribuição (Magalhães & Lima, 2008).

3. DESENVOLVIMENTO DOS TESTES DE HIPÓTESES PARA A DISTRIBUIÇÃO EXPONENCIAL

Mediante ao exposto, o objetivo é apresentar como os dois testes, convencional e sequencial, são construídos para o parâmetro da distribuição exponencial.

3.1 TESTE DE HIPÓTESES CONVENCIONAL PARA A DISTRIBUIÇÃO EXPONENCIAL

Para encontrar a função densidade de probabilidade da estatística $B(X)$, razão de verossimilhanças e sua distribuição acumulada de uma amostra advinda de uma distribuição Exponencial, foi desenvolvida a teoria de teste de hipóteses para esta distribuição de acordo com Brighenti (2007).

Seja $X_1,\ldots,X_n$ uma amostra aleatória de uma distribuição exponencial $f(x;\theta)=\theta e^{-\theta x}$. Um teste mais poderoso de tamanho α para testar as hipóteses simples $H_0:\theta=\theta_0$ versus $H_1:\theta=\theta_1$, sendo $\theta_0<\theta_1$, é o de razão de verossimilhanças, portanto de (2) tem-se a função de verossimilhança para a distribuição exponencial sob a hipótese nula:

$$L_0=\prod_{i=1}^{n} f(x_i;\theta_0)=\prod_{i=1}^{n}\theta_0 e^{-\theta_0 x_i}=(\theta_0 e^{-\theta_0 x_1})\cdot\ldots\cdot(\theta_0 e^{-\theta_0 x_n})=\theta_0{}^n e^{-\theta_0(x_1+x_2+\ldots+x_n)}=\theta_0^n e^{-\theta_0\sum_{i=1}^{n} x_i} \quad (14)$$

De modo análogo, obtém-se L_1. Portanto, tem-se:

$$L_0=\theta_0^n \exp\left\{-\theta_0\sum_{i=1}^{n} x_i\right\} \quad \text{e} \quad L_1=\theta_1^n \exp\left\{-\theta_1\sum_{i=1}^{n} x_i\right\} \quad (15)$$

E a razão de verossimilhanças, dada por (3), será:

$$B(x_1,\ldots,x_n)=\frac{L_0}{L_1}=\frac{\theta_0^n \exp\left\{-\theta_0\sum_{i=1}^{n} x_i\right\}}{\theta_1^n \exp\left\{-\theta_1\sum_{i=1}^{n} x_i\right\}}=\left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{-(\theta_0-\theta_1)\sum_{i=1}^{n} x_i\right\}. \quad (16)$$

A família de testes dada pela razão de verossimilhanças consiste em rejeitar H_0 se:

$$B(x_1,\ldots,x_n)\le b_c \Rightarrow \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{-(\theta_0-\theta_1)\sum_{i=1}^{n} x_i\right\}\le b_c \quad (17)$$

Para obter, entre esses testes, aquele de tamanho pré-determinado, por exemplo, com $\alpha = 0,05$, o procedimento usual é obter a desigualdade equivalente, de acordo com (5):

$$\left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{-(\theta_0-\theta_1)\sum_{i=1}^n x_i\right\} \le \mathrm{b}_c \Rightarrow \exp\left\{-(\theta_0-\theta_1)\sum_{i=1}^n x_i\right\} \le \left[\mathrm{b}_c\left(\frac{\theta_1}{\theta_0}\right)^n\right] \Rightarrow$$

$$\Rightarrow \ln \mathrm{e}^{\left\{-(\theta_0-\theta_1)\sum_{i=1}^n x_i\right\}} \le \ln\left[\mathrm{b}_c\left(\frac{\theta_1}{\theta_0}\right)^n\right] \Rightarrow -(\theta_0-\theta_1)\sum_{i=1}^n x_i \le \ln\left[\mathrm{b}_c\left(\frac{\theta_1}{\theta_0}\right)^n\right] \Rightarrow$$

$$\Rightarrow \sum_{i=1}^n x_i \le \frac{1}{(\theta_1-\theta_0)} \ln\left[b_c\left(\frac{\theta_1}{\theta_0}\right)^n\right] = k \qquad (18)$$

Logo, $\sum_{i=1}^n x_i \le \frac{1}{(\theta_1-\theta_0)} \ln\left[\mathrm{b}_c\left(\frac{\theta_1}{\theta_0}\right)^n\right]$ é a desigualdade equivalente em termos de uma estatística suficiente S, e esta independe dos parâmetros.

Portanto, rejeita-se H_0 se $\sum_{i=1}^n x_i \le \frac{1}{(\theta_1-\theta_0)} \ln\left[\mathrm{b}_c\left(\frac{\theta_1}{\theta_0}\right)^n\right]$.

Tem-se que $Y = \sum_{i=1}^n x_i$ possui uma distribuição Gama com parâmetros n e θ. Isto é, $f_{\sum_{i=1}^n x_i}(y) = \frac{1}{\Gamma(n)} y^{n-1}\theta^n e^{-\theta y}$. Assim, para obter o valor de k, resolve-se a equação integral:

$$P_{\theta_0}\left[\sum_{i=1}^n x_i \le k\right] = \int_0^k \frac{1}{\Gamma(n)} y^{n-1}\theta_0^n e^{-\theta_0 y} dy = 0,05\,. \qquad (19)$$

Para calcular diretamente a distribuição da variável $B(X_1,...,X_n) = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{-(\theta_0-\theta_1)\sum_{i=1}^n x_i\right\}$, deve-se determinar a densidade de B, através do teorema de transformação de variáveis: $f_B(b) = \left|\frac{d}{db} g^{-1}(b)\right| f_Y\left(g^{-1}(b)\right)$ (Mood et al., 1974).

Tem-se que $B = g(\mathrm{Y}) = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\{-(\theta_0-\theta_1)Y\}$ é a exponencial de uma variável com distribuição Gama de parâmetros (n,θ). Expressando Y como função de B, tem-se que:

$$g^{-1}(b)=\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\Rightarrow\frac{d}{db}g^{-1}(b)=0.\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]+\frac{1}{(\theta_1-\theta_0)}\cdot\frac{1}{b}=\frac{1}{(\theta_1-\theta_0)\mathrm{b}}$$

para $\theta_1>\theta_0$. (20)

$$f_Y\left(g^{-1}(b)\right)=\frac{\theta^n}{\Gamma(n)}\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}^{n-1}e^{-\theta\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}} \quad (21)$$

Utilizando (20) e (21), tem-se:

$$f_B(b)=\left|\frac{d}{db}g^{-1}(b)\right|f_Y\left(g^{-1}(b)\right)=\frac{1}{(\theta_1-\theta_0)b}\frac{\theta^n}{\Gamma(n)}\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}^{n-1}e^{-\theta\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}}$$

$$f_B(b)=\frac{1}{(\theta_1-\theta_0)b}\frac{\theta^n}{\Gamma(n)}\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}^{n-1}e^{-\theta\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}} \quad (22)$$

Para um determinado valor de α, é necessário resolver a equação integral:

$$\int_0^{b_c}\frac{1}{(\theta_1-\theta_0)b}\frac{\theta^n}{\Gamma(n)}\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}^{n-1}e^{-\theta\left\{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right\}}db=\alpha$$

Usando substituição de variáveis tem-se:

$$F_B^0(b)=\int_0^{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]}\frac{\theta^n}{\Gamma(n)}u^{n-1}e^{-\theta u}du$$

E para $\theta=\theta_0$, tem-se:

$$F_B^0(b)=\int_0^{\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]}\frac{{\theta_0}^n}{\Gamma(n)}u^{n-1}e^{-\theta_0 u}du \quad (23)$$

Fazendo novamente uma substituição $w=\theta_0 u$:

$$u=\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\Rightarrow w=\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]$$

$$F_B^0(b)=\int_0^{\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]}\frac{1}{\Gamma(n)}w^{n-1}e^{-w}dw=\Gamma\left(n,\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right),$$

em que $\Gamma(n,t)=\left(\frac{1}{\Gamma(n)}\right)\int_0^t y^{n-1}\exp\{-y\}dy$ é a função gama incompleta.

Então, tem-se:

$$F_B^0(b)=P_{\theta=\theta_0}\left(B(X)\le b\right)=\Gamma\left(n,\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right) \tag{24}$$

De maneira análoga, obtém-se para $\theta=\theta_1$, fazendo uma substituição $w=\theta_1 u$ em (23):

$$u=\frac{1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\Rightarrow w=\frac{\theta_1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]$$

$$F_B^1(b)=\int_0^{\frac{\theta_1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]}\frac{1}{\Gamma(n)}w^{n-1}e^{-w}dw \tag{25}$$

$$F_B^1\left(b\right)=P_{\theta=\theta_1}\left(B(X)\le b\right)=\Gamma\left(n,\frac{\theta_1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b\right]\right) \tag{26}$$

Para $\alpha=0,05$, tem-se a equação integral:

$$F_B^0(b)=P_{\theta=\theta_0}\left(B(X)\le b_c\right)=\Gamma\left(n,\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b_c\right]\right)=0,05$$

Uma maneira de garantir a importância de ambos os erros é obter testes em que $\alpha=\beta$, isto é:

$$F_B^0\left(b_c\right)=1-F_B^1\left(b_c\right) \tag{27}$$

Neste caso, garante-se que se um deles for pequeno, o outro também será e não há um valor pré-fixado para α e sim uma condição $\alpha=\beta$ que fornece um determinado valor crítico b_c.

Para garantir $\alpha=\beta$, equivale a encontrar o valor de b_c que satisfaça a seguinte equação:

$$\Gamma\left(n,\frac{\theta_0}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b_c\right]\right)=1-\Gamma\left(n,\frac{\theta_1}{(\theta_1-\theta_0)}\ln\left[\left(\frac{\theta_1}{\theta_0}\right)^n b_c\right]\right) \tag{28}$$

Portanto, o valor de b_c corresponde ao valor crítico na função densidade de probabilidade B(X) da estatística de razão de verossimilhança da distribuição Exponencial conforme inequação (17).

3.2 TESTE DE HIPÓTESES SEQUENCIAL PARA A DISTRIBUIÇÃO EXPONENCIAL

Para a realização do teste de hipóteses sequencial obtém-se o plano de decisão sequencial no qual são definidas três regiões: a região de aceitação da hipótese nula, de rejeição da hipótese nula e de continuação do teste. Essas regiões podem ser construídas através de retas e com auxílio de um procedimento gráfico pode-se tomar decisões (Estefanel & Barbin, 1979).

A obtenção deste plano para a distribuição exponencial baseado em Wald (1947) é através do cálculo da soma cumulativa do logaritmo da razão de verossimilhança, $\log \Lambda_i$, à medida que novos dados são avaliados, com $S_0 = 0$, para $i = 1, 2, \ldots$ Desse modo, $S_i = S_{i-1} + \log \Lambda_i$.

Em que $\log \Lambda_i$, para uma amostra, é:

$$\log \Lambda(x) = \log\left(\frac{\theta_1^{-1} e^{-\frac{x}{\theta_1}}}{\theta_0^{-1} e^{-\frac{x}{\theta_0}}}\right) = \log\left(\frac{\theta_0}{\theta_1} e^{\frac{x}{\theta_0} - \frac{x}{\theta_1}}\right) = -\log\left(\frac{\theta_1}{\theta_0}\right) + \left(\frac{\theta_1 - \theta_0}{\theta_0 \theta_1}\right) x \tag{29}$$

A soma acumulada para todo x é:

$$S_n = \sum_{i=1}^{n} \log \Lambda(x_i) = -n \log\left(\frac{\theta_1}{\theta_0}\right) + \left(\frac{\theta_1 - \theta_0}{\theta_0 \theta_1}\right) \sum_{i=1}^{n} x_i \tag{30}$$

A regra de parada será:

- Se $S_i \leq A$, não se rejeita H_0;
- Se $S_i \geq B$, rejeita-se H_0;
- Se $A < S_i < B$, continua-se o experimento aumentando a amostra.

Assim,

$$A < -n \log\left(\frac{\theta_1}{\theta_0}\right) + \left(\frac{\theta_1 - \theta_0}{\theta_0 \theta_1}\right) \sum_{i=1}^{n} x_i < B \tag{31}$$

Reorganizando (31) para facilitar, tem-se:

$$\begin{aligned} &A + n \log\left(\frac{\theta_1}{\theta_0}\right) < \left(\frac{\theta_1 - \theta_0}{\theta_0 \theta_1}\right) \sum_{i=1}^{n} x_i < B + n \log\left(\frac{\theta_1}{\theta_0}\right) \Rightarrow \\ &\Rightarrow \left[A + n \log\left(\frac{\theta_1}{\theta_0}\right)\right] \cdot \left(\frac{\theta_0 \theta_1}{\theta_1 - \theta_0}\right) < \sum_{i=1}^{n} x_i < \left[B + n \log\left(\frac{\theta_1}{\theta_0}\right)\right] \cdot \left(\frac{\theta_0 \theta_1}{\theta_1 - \theta_0}\right) \end{aligned} \tag{32}$$

Portanto, tem-se as seguintes regras para tomar uma decisão:

- Se $\left[A + n \log\left(\frac{\theta_1}{\theta_0}\right)\right] \cdot \left(\frac{\theta_0 \theta_1}{\theta_1 - \theta_0}\right) < \sum_{i=1}^{n} x_i$, não se rejeita H_0;

- Se $\left[B+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)>\sum_{i=1}^{n}x_i$, rejeita-se H_0;
- Se $\left[A+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)<\sum_{i=1}^{n}x_i<\left[B+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$,

continua-se o experimento.

Desse modo, pode-se utilizar um procedimento gráfico, onde $\left[A+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$ e $\left[B+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$, determinam retas paralelas com coeficiente angular $\log\left(\frac{\theta_1}{\theta_0}\right)\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$ e vão ser denotadas por: $h_0=\left[A+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$ e $h_1=\left[B+n\log\left(\frac{\theta_1}{\theta_0}\right)\right]\cdot\left(\frac{\theta_0\theta_1}{\theta_1-\theta_0}\right)$. Elas são consideradas limites pois se o ponto indicado por $\sum_{i=1}^{n}x_i$, cair abaixo de h_0, aceita-se a hipótese nula, se o ponto cair acima da reta superior, rejeita-se a hipótese nula. Se o ponto ficar entre as duas linhas paralelas, a avaliação deve ser continuada sem nenhuma decisão (Nagelkerke & Hart, 1980).

4. APLICAÇÃO

Para aplicar o teste de hipóteses convencional e sequencial para a distribuição Exponencial, considerou-se, a seguir, um exemplo hipotético.

Exemplo:

Uma fábrica utiliza dois métodos para a produção de lâmpadas, método A e método B, tendo eles diferença de custo de produção e durabilidade. O tempo de duração das lâmpadas produzidas pelo método A seguem uma distribuição Exponencial com parâmetro 1/100 e o tempo de duração das lâmpadas do método B, seguem uma Exponencial de parâmetro 1/200. Ao realizar a compra de lâmpadas em uma loja, suponha que, para o consumidor final não haja diferença entre as embalagens. Portanto, este não consegue distinguir se está comprando lâmpadas produzidas pelo método A ou B. Para verificar por qual método a lâmpada foi produzida, é necessário que o consumidor registre o tempo de duração de cada lâmpada.

Supondo a compra de 20 lâmpadas e o registro de tempo de duração destas, é possível realizar o teste da seguinte maneira:

H_0 = as lâmpadas comercializadas pela loja são provenientes do método B de produção;

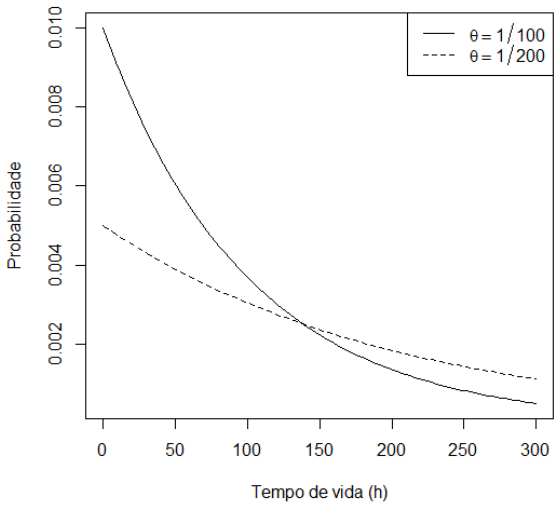
H_1 = as lâmpadas comercializadas pela loja são provenientes do método A de produção.

Utilizando o comando *rexp(20,0.01)* no software R (R Core Team, 2019), pode-se gerar uma amostra de valores de uma distribuição exponencial, em que o valor 20 é o tamanho da amostra e 0,01 o parâmetro taxa da Exponencial. Assim, considerando o método A de produção, tem-se os seguintes dados simulados de durabilidade, em horas:

lamp=c(110.18,94.86,23.62,329.31,22.06,32.78,9.33,5.55,187.23,159.66,119.41,6.11,15.92,42,3.46,283.20,36.93,31.03,106.18,62.11)

O erro tipo I foi considerado igual a 5%, assim $\alpha = 0,05$. A Figura 1 mostra a função densidade de probabilidade para as distribuições com os diferentes parâmetros.

Figura 1 - Função densidade de Probabilidade para X~Exp (1/100) e X~Exp (1/200)



Fonte – Autoras

4.1TESTE DE HIPÓTESES CONVENCIONAL

Primeiramente, para o teste de hipóteses convencional tem-se o seguinte desenvolvimento:

Sob as hipóteses:

$$\begin{cases} H_0 : \theta = \dfrac{1}{200} \text{ (Método B)} \\ H_1 : \theta = \dfrac{1}{100} \text{ (Método A)} \end{cases} \text{, já que } \theta_1 > \theta_0 \qquad (33)$$

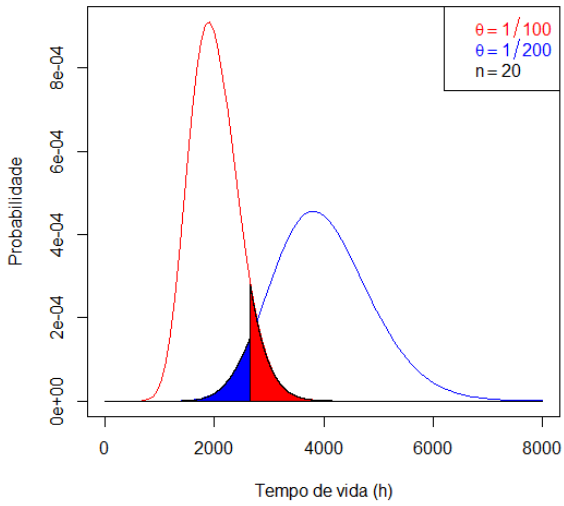
Para uma amostra com distribuição Gama e parâmetros $n=20$ e $\theta=\frac{1}{100}$, obteve-se $\sum_{i=1}^{20} x_i = 1680{,}93$. Pela expressão (19) e utilizando a estatística suficiente, encontra-se o valor crítico através de:

$$P_{\theta_0}\left[1680{,}93 \leq k\right] = \int_0^k \frac{1}{\Gamma(20)} y^{19} \left(\frac{1}{200}\right)^{20} e^{-\frac{y}{200}} dy = 0{,}05 \qquad (34)$$

Utilizando o software R (R Core Team, 2019), como auxílio para calcular a integral acima, obteve-se que 2650,83 corresponde ao valor crítico, logo, *k* = 2650,93, e a integral corresponde a 0,0001522128.

Assim, gerou-se a Figura 2, com as áreas hachuradas correspondentes ao erro tipo I, $\alpha = 0{,}05$, da distribuição sob $H_0 : \theta = \frac{1}{200}$ e ao erro tipo II, β, da distribuição sob $H_1 : \theta = \frac{1}{100}$, a partir do software R (R Core Team, 2019).

Figura 2 - Áreas correspondentes às probabilidades de erros tipos I e II para o caso de $n=20$



Fonte - Autoras

Assim, o critério estabelecido anteriormente por (18) é: rejeita-se H_0 se $\sum x_i \leq k$, desse modo como *k*=2650,93 e $\sum_{i=1}^{20} x_i = 1680{,}93$, logo, conclui-se que as lâmpadas comercializadas pela loja são provenientes do método A de produção.

Neste teste considerou-se o tamanho *n* fixo em 20, e $\alpha = 0{,}05$, no entanto β não foi definido. Ele deve ser calculado no ponto *k* a partir da distribuição sob H_1, ou seja, $\beta = 1 - \int_0^{2650{,}93} \frac{1}{\Gamma(20)} y^{19} \left(\frac{1}{100}\right)^{20} e^{-\frac{y}{100}} dy$, e utilizando novamente o auxílio do software R (R Core Team, 2019) para calcular β, tem-se, $\beta = 0{,}0815$, que corresponde a um poder do teste de $1-\beta = 0{,}9185$.

Assim, para realizar o teste de hipóteses convencional com erros tipo I e tipo II iguais, com a finalidade de manter ambos os erros pequenos, seria necessário aplicar a fórmula (28) desenvolvida anteriormente:

$$\Gamma\left(20,\frac{0,005}{(0,01-0,005)}\ln\left[\left(\frac{0,01}{0,005}\right)^{20}b_c\right]\right)=1-\Gamma\left(20,\frac{0,01}{(0,01-0,005)}\ln\left[\left(\frac{0,01}{0,005}\right)^{20}b_c\right]\right) \quad (35)$$

Se b_c for calculado por (18), que corresponde a um $\alpha=0,05$. Tem-se:

$$\frac{1}{(0,01-0,005)}\ln\left[b_c\left(\frac{0,01}{0,005}\right)^{20}\right]=2650,93\Rightarrow e^{13,25465}=1048576b_c\Rightarrow b_c=0,5443 \quad (36)$$

No entanto,

$$\Gamma\left(20,\frac{0,005}{(0,01-0,005)}\ln\left[\left(\frac{0,01}{0,005}\right)^{20}0,5443\right]\right)\neq$$
$$\neq 1-\Gamma\left(20,\frac{0,01}{(0,01-0,005)}\ln\left[\left(\frac{0,01}{0,005}\right)^{20}0,5443\right]\right) \quad (37)$$

Assim,

$$\int_0^{2650,93}\frac{1}{\Gamma(20)}y^{19}\left(\frac{1}{200}\right)^{20}e^{-\frac{y}{200}}dy \text{ e } 1-\int_0^{2650,93}\frac{1}{\Gamma(20)}y^{19}\left(\frac{1}{100}\right)^{20}e^{-\frac{y}{100}}dy \quad (38)$$

Correspondem a áreas diferentes, que calculadas pelo R (R Core Team, 2019) fornecem as seguintes probabilidades: 0,05000114 e 0,0815157.

Ou seja, não é possível tornar fixo os erros tipo I, tipo II e o tamanho da amostra. Logo, para encontrar áreas iguais correspondentes aos erros tipo I e tipo II, iguais a 5%, deve-se aumentar o tamanho da amostra.

4.2TESTE DE HIPÓTESES SEQUENCIAL

Para o teste de hipóteses sequencial considerou-se a taxa de falha $1/\theta$, e consequentemente o parâmetro de escala θ, logo o teste de hipóteses foi formulado por:

$$\begin{cases}H_0:\theta=100 \text{ (Método A)}\\ H_1:\theta=200 \text{ (Método B)}\end{cases}, \text{ já que } \theta_1>\theta_0 \quad (39)$$

Considerou-se os erros iguais a 5%, para mantê-los em uma probabilidade pequena. Logo, ao calcular os limites A e B tem-se, que $\mathrm{A}=\log\frac{\beta}{(1-\alpha)}=\log\frac{0,05}{0,95}=-2,944439$ e $\mathrm{B}=\log\frac{(1-\beta)}{\alpha}=\log\frac{0,95}{0,05}=2,944439$, substituindo A e B na fórmula desenvolvida em (32), tem-se:

$$A + n\log\left(\frac{\theta_1}{\theta_0}\right) < + \left(\frac{\theta_1 - \theta_0}{\theta_0\theta_1}\right)\sum_{i=1}^{n} x_i < B + n\log\left(\frac{\theta_1}{\theta_0}\right) \Rightarrow$$

$$\Rightarrow -2,944439 + n\log 2 < \left(\frac{200-100}{200\cdot 100}\right)\sum_{i=1}^{n} x_i < 2,944439 + n\log 2 \Rightarrow \quad (40)$$

$$\Rightarrow -588,8878 + 138,6294n < \sum_{i=1}^{n} x_i < 588,8878 + 138,6294n$$

Portanto, $-588,8878 + 138,6294n = h_0$ e será a reta de não rejeição da hipótese nula e $588,8878 + 138,6294n = h_1$ de rejeição.

Para cada valor de n são apresentados os limites na Tabela 1 a seguir.

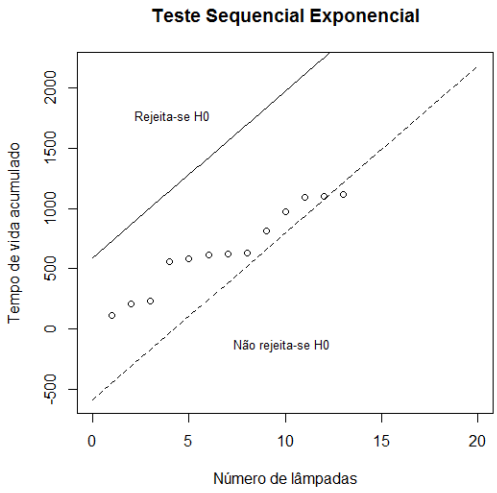
Tabela 1 - Valores para os diferentes tamanhos de amostras

n	x_i	$\sum_{i=1}^{n} x_i$	h_0	h_1
1	110,18	110,18	-450	728
2	94,86	205,04	-311	866
3	23,62	228,66	-173	1.005
4	329,31	557,97	-34	1.143
5	22,06	580,03	104	1.282
6	32,78	612,81	243	1.421
7	9,33	622,14	381	1.559
8	5,55	627,69	520	1.697
9	187,23	814,92	659	1.836
10	159,66	974,58	797	1.975
11	119,41	1.093,99	936	2.114
12	6,11	1.100,10	1.075	2.252
13	15,92	1.116,02	1.213	2.391
14	42,00	1.158,02	1.352	2.530
15	3,46	1.161,48	1.490	2.668
16	283,20	1.444,68	1.629	2.807
17	36,93	1.481,61	1.768	2.945
18	31,03	1.512,64	1.906	3.084
19	106,18	1.618,82	2.045	3.223
20	62,11	1.680,93	2.184	3.361

Fonte - Autoras

A partir do software R (R Core Team, 2019) foi criado um script para construção do gráfico na Figura 3:

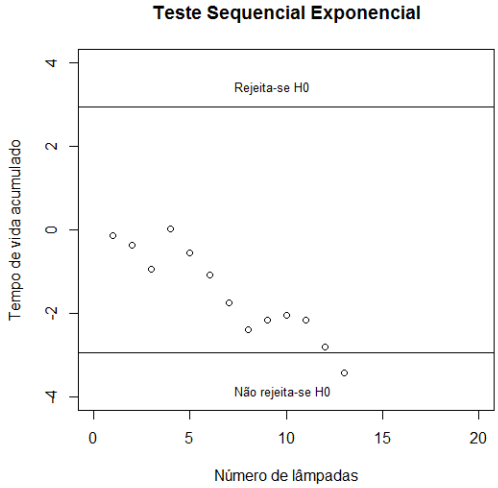
Figura 3 - Teste sequencial da razão de verossimilhanças da distribuição exponencial



Fonte - Autoras

Assim, pelo gráfico é possível observar que não se rejeita a hipótese nula com n=13, pois o ponto ficou abaixo de h_o, ainda pode-se verificar pela Tabela 1 que para n=13, $\sum_{i=1}^{n} x_i = 1116,02$, logo $1116,02 < 1213,295$, concluindo que as lâmpadas comercializadas pela loja são provenientes do método A de produção. Além disso, foi gerado o gráfico para o logaritmo da razão de verossimilhança, apresentado na Figura 4:

Figura 4 - Teste sequencial do logaritmo da razão de verossimilhanças da distribuição exponencial



Fonte - Autoras

É possível observar que neste caso, o teste sequencial conseguiu manter os erros com uma probabilidade pequena, ambos iguais a 5%, e também o tamanho da amostra menor, correspondente a 13 lâmpadas. Então, conseguiu-se tomar a mesma decisão do teste de hipóteses convencional executado anteriormente, entretanto sem a necessidade de testar todas as 20 lâmpadas, otimizando assim o tempo de teste.

5.CONCLUSÃO

No exemplo hipotético realizado neste artigo, considerando os testes de hipóteses convencional e sequencial para o parâmetro da distribuição Exponencial, pode-se concluir que, no caso sequencial foi possível a tomada de decisão com um tamanho de amostra menor quando comparada a convencional, diminuindo o tempo e custo para a execução do teste.

Utilizando os conceitos desenvolvidos do Teste de Hipóteses Convencional e Sequencial sobre o parâmetro da distribuição Exponencial é possível realizar as mesmas análises variando os parâmetros para outras comparações.

6.REFERÊNCIAS BIBLIOGRÁFICAS

Barbosa, J. C. *A amostragem seqüencial*, In: Fernandes, O. A.; Correia, A. C. B.; De Bortoli, S. A. (ed.). (1992). *Manejo integrado de pragas e nematóides.* Jaboticabal: FUNEP, 205-211.

Brighenti, C. R. G. (2007). Testes frequentistas condicionais e testes com interpretação bayesiana e frequentista condicional. 212f. (Tese de doutorado). Universidade Federal de Lavras (UFLA), Lavras.

Brighenti, C. R. G., Resende, M. & Brighenti, D. M. (2013). Estimação sequencial bayesiana do parâmetro da distribuição Poisson e sua aplicação na modelagem da infestação de galhas de Psilídeo em alecrim do campo. Revista Brasileira de Estatística, 74(238), 129-151.

Bussab, W. O. & Morettin, P. A. (2017). *Estatística Básica* (9a ed.). São Paulo: Saraiva.

Caetano, S. L. & Louzada-Neto, F. (2007). Controle de qualidade via dados acelerados com distribuição exponencial e relação estresse-resposta lei de potência inversa. Revista de Matemática e Estatística, 25(1), 85-98.

Casella, G. & Berger R. L. (2002) *Statistical Inference* (2a ed.). CA: Duxbury.

Estefanel, V. & Barbin, D. (1979). Amostragem sequencial baseada no teste sequencial da razão de probabilidades e seu uso na determinação da época de controle das lagartas da soja no estado do Rio Grande do Sul. Revista do Centro Ciência Rurais, 9(1), 29-48.

Fernandes, M. G., Busoli, A. C. & Barbosa, J. C. (2002). Amostragem seqüencial de Spodoptera frugiperda (J. E. Smith, 1797) (lepidoptera, noctuidae) em algodoeiro. Revista Brasileira Agrociência, 8(3), 213-218.

Fonseca, D. R. (2008). Estimador de Máxima Verossimilhança utilizado em Testes de Vida Sequenciais com Trucagem: uma aplicação com um modelo Weibull de três

parâmetros. (Dissertação de mestrado). Universidade Estadual do Norte Fluminense (UENF), Campos dos Goytacazes - Rio de Janeiro.

Leal, V. J. & Andrade, P. C. de R. (2018). Modelagem dos dados de falha de um caminhão fora de estrada. ForScience, 6(3).

Liu, Y., Chen, Y. & Cheng, J. (2009). A comparative study of optimization methods and conventional methods for sampling design in fishery-independent surveys. ICES Journal of Marine Science, storebo, 66, 1873-1882.

Macedo, C. C. L. (2018). Análise de confiabilidade baseada em ensaios acelerados de vida: estudo de caso de lâmpadas incandescentes e LED utilizadas em refrigeradores. 132 f. (Trabalho de Conclusão de Curso). Universidade Tecnológica Federal do Paraná, Curitiba.

Magalhães, M.N. & Lima, A.C.P. (2008). *Noções de Probabilidade e Estatística* (6a ed.). São Paulo: USP.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974). *Introduction to the theory of statistics* (3a ed.). Singapore: McGraw-Hill International.

Nagelkerke, N. J. D. & Hart, A. A. M. (1980). The sequential comparison of survival curves. Biometrika, 67(1), 247-249.

Penteado, S. R. C., De Oliveira, E. B. & Iede, E. T. (2008). Utilização da amostragem seqüencial para avaliar a eficiência do parasitismo de Deladenus (Beddingia) siricidicola (Nematoda: Neotylenchidae) em adultos de Sirex noctilio (Hymenoptera: Siricidae). Ciência Florestal, 18(2), 223-231.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Áustria.

Wald, A. (1947). *Sequential Analysis.* New York: John Willey & Sons, Inc., USA.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

COMPORTAMENTO TEMPORAL DO ÍNDICE DE EXPANSÃO DO COMÉRCIO (IEC) NA CIDADE DE SÃO PAULO

Fábio Henrique Florindo Amano

henriquefabio205@gmail.com

Universidade Federal de São João del Rei-UFSJ

Bathyelly Juncal Alves Batista

bathyelly@hotmail.com

Universidade Federal de São João del Rei-UFSJ

Luciane Teixeira Passos Giarola

luciane@ufsj.edu.br

Universidade Federal de São João del Rei-UFSJ

Rejane Corrêa da Rocha

rejane@ufsj.edu.br

Universidade Federal de São João del Rei-UFSJ

Resumo: O índice de expansão do comércio avalia a expectativa dos empresários ao que se refere à ampliação de seus negócios, sendo afetado por mudanças na economia. Por ser muito específico, são necessários estudos que avaliem o seu comportamento e possam estimar valores futuros, a fim de auxiliar decisões governamentais e empresariais. Em virtude disso, este trabalho se propôs a analisar mensalmente esse índice, no período de junho de 2011 a julho de 2020, seguindo a metodologia proposta por Box e Jenkins em 1976, e fazer previsões futuras. Utilizando o critério de Akaike foram selecionados modelos cujos resíduos apresentaram características de ruído branco. Por razão da identificação de presença da componente de tendência foram utilizados modelos integrados. O indicador de erro percentual absoluto médio (MAPE) foi utilizado para escolha do melhor modelo, isto é, o que gera as previsões mais acuradas. O modelo ARIMA (2,1,0) foi o que apresentou o menor MAPE no valor aproximado de 6% e foi capaz de captar a dinâmica temporal da série.
Palavras-chave: modelagem estatística; modelo ARIMA; previsão; IEC.

Abstract: The trade expansion index assesses the expectations of entrepreneurs regarding the expansion of their businesses, being affected by changes in the economy. Because it is very specific, studies are needed to assess their behavior and to estimate future values in order to assist government and business decisions. As a result, this work proposed to analyze this index monthly, from June 2011 to July 2020, following the methodology proposed by Box and Jenkins in 1976, and to make future predictions. Using the Akaike criterion, models were selected whose

residues had white noise characteristics. Due to the identification of the presence of the trend component, integrated models were used. The average absolute percentage error indicator (MAPE) was used to choose the best model, that is, the one that generates the most accurate forecasts. The ARIMA model (2,1,0) was the one that presented the lowest MAPE in the approximate value of 6% and was able to capture the time dynamics of the series.
Keywords: statistical modeling; ARIMA model; forecast; TEI.

1. INTRODUÇÃO

A cidade de São Paulo é responsável por alocar aproximadamente 1/3 do comércio brasileiro além de registrar uma receita anual bruta de mais de R$1 trilhão. Sua região metropolitana atua como o maior polo de riqueza nacional, tendo apresentado em 2017 um Produto Interno Bruto (PIB) referente a aproximadamente 10,6% do PIB brasileiro (Rodrigues, 2019). Em 2020, de acordo com o Instituto Brasileiro de Geografia Estatística (IBGE) e a Fundação Sistema Estadual de Análise de Dados (Seade) essa participação foi de 31,2% (Investsp, 2021). Isso evidencia o quanto o estado influencia na economia brasileira. A principal atividade econômica da capital é o varejo, seguido pelo atacado e a comercialização de veículos, peças e motocicletas (Investsp, 2020).

No estado de São Paulo, foi criada, em 1938, a Federação do Comércio com o objetivo de representar os interesses dos comerciantes e empresas do setor, além de contribuir para sua modernização. Nos dias atuais recebe o nome de Federação do Comércio de Bens, Serviços e Turismo do Estado de São Paulo (FecomercioSP) e possui atuação voltada para a promoção do crescimento econômico do Brasil. Essa entidade busca um diálogo entre capital e trabalho, defendendo o mercado interno, a livre iniciativa, a desestatização e o tratamento diferenciado para as pequenas e microempresas (Fecomerciosp, 2020).

Adicionalmente, a FecomercioSP desenvolve análises em diversos campos relevantes para determinado setor ou empresa, promovendo estudos que auxiliam tanto o poder público como outras entidades que necessitem desse tipo de material. Aqui, pode-se citar algumas delas: o Custo de Vida por Classe Social (CVCS), Índice de Preços de Serviços (IPS), Pesquisa de Endividamento e Inadimplência das Famílias (PEIC), o Índice de Confiança do Consumidor (ICC), Índice de Confiança do Empresário do Comércio (ICEC) e o Índice de Expansão do Comércio (IEC), tema deste artigo. Ainda, promove debates sobre pautas importantes, que afetam a vida de empresários do comércio de bens, serviços e turismo; defende ações que trazem bem-estar social e combate práticas que colocam em risco a estrutura empresarial paulista (Fecomerciosp, 2020).

No que se refere ao IEC, a FecomercioSP realiza sua apuração mensalmente e sua variação vai de zero a duzentos pontos, com tais extremos representando, respectivamente, o desinteresse e o interesse absoluto na expansão dos negócios. Ele gera resultados tempestivos, que acompanham as principais mudanças que ocorrem na economia, principalmente na área de negócios. O IEC é obtido por meio de entrevistas com cerca de 600 empresários

do comércio dos municípios que compõem a Região Metropolitana de São Paulo (FECOMERCIO, 2020).

O IEC auxilia a interpretar constantes mudanças que ocorrem na economia, muitas afetando os investimentos no setor empresarial. Sua análise identifica a perspectiva dos empresários do comércio em relação a contratações, a compra de máquinas ou equipamentos e a abertura de novas lojas (Fernandes, 2021).

Segundo a pesquisa desenvolvida pela FecomercioSP, o IEC do município passou de 101,6 pontos em abril de 2018 para 102,6 em maio do mesmo ano, registrando uma alta de 1% e sendo, até esse período, a maior pontuação desde dezembro de 2014. O nível de investimento das empresas também apresentou alta de 1% na comparação mensal do mesmo período. Posteriormente a abril de 2018, o índice experimentou oscilações e apresentou crescimento no segundo semestre de 2019, atingindo novo marco de 113,46 pontos em dezembro. O ano de 2020 iniciou-se com sentimento de otimismo por parte dos empresários, com consumo aquecido e expectativa de expansão das vendas do comércio, em parte devido à melhora dos indicadores macroeconômicos em 2019 e às medidas governamentais (Oliveira, 2020).

Porém, a partir do ano de 2020, a economia brasileira sofreu o impacto da pandemia da Covid 19, em particular a indústria e o comércio. O primeiro caso de COVID- 19 no Brasil ocorreu no estado de São Paulo no dia 26 de fevereiro. Diante do número crescente de mortos, a Organização Mundial da Saúde (OMS) decretou situação de emergência de saúde internacional e no Brasil foi decretada a quarentena a partir de março de 2020 (DECRETO Nº 64.881 de 22/3/2020) (Aquino et al, 2020). Assim, a fim de conter o avanço da doença, foram impostos o isolamento social e as restrições no funcionamento das atividades comerciais. Além disso, o comércio de São Paulo experimentou uma escassez de produtos importados, principalmente, vindos da China, país que registrou as primeiras contaminações pelo vírus (Zu et al, 2020).

Assim, o IEC começou a cair a partir de janeiro de 2020, se intensificando nos meses seguintes. Uma queda de 3% foi observada de março para abril. Em maio o índice caiu mais, passando de 107,03 pontos do mês de abril para 87,53 pontos. No auge das medidas de restrição e circulação de pessoas, a redução foi ainda maior no mês de junho, declinando para 62,77 pontos, 28% menor que no mês anterior. Também o ICEC atingiu neste mesmo mês o baixo valor de 61 pontos (Fernandes, 2021). No mesmo período de 2019, esse índice foi de 117 pontos. Isto mostra como a confiança dos empresários foi

afetada pelas turbulências e incertezas trazidas pela pandemia, muito embora o ano de 2020 tenha se inaugurado sob uma esfera otimista.

Diante disso, fica evidente a importância de um estudo e acompanhamento da evolução do IEC. Além disso, observa-se que esse índice, por ser muito específico, ainda carece de estudo sobre seu comportamento. Então, neste trabalho pretendeu-se estudar o comportamento da série histórica desse índice, buscando captar características tais como tendência e sazonalidade. Pretendeu-se ainda, obter um modelo de previsão que possa ser utilizado para fornecer estimativas futuras desse índice. As previsões poderão auxiliar o Poder Público bem como quaisquer outras entidades que necessitem desta informação, na medida em que o índice identifica a percepção dos empresários sobre as intenções de expansão das atividades comerciais.

Para tanto, foi utilizada a metodologia de Box & Jenkins (1976), em virtude da relação existente entre o IEC e o domínio do tempo. Essa metodologia mostrou-se eficiente nas projeções feitas por Silva et al (2020) para o PIB brasileiro nos anos de 2018 e 2019 quando comparadas com projeções oficiais. Oikawa (2020) também utilizou os modelos de Box e Jenkins com intervenção para avaliar a influência da crise financeira americana ocorrida em 2008 na produção física industrial brasileira e concluiu que houve impacto negativo com recuperação econômica gradual, mas relevante.

2. ANÁLISE DA SÉRIE TEMPORAL DO IEC DE SÃO PAULO

Os dados foram obtidos no site da Federação de Comércio de São Paulo (FecomercioSP) e referem-se ao Índice de Expansão do Comércio (IEC), apurado mensalmente. Trata-se de um *score* baseado nas perspectivas futuras dos empresários. Tais dados são coletados, periodicamente, desde 2011, totalizando 113 observações mensais, de junho de 2011 a julho de 2020.

O processo de análise iniciou-se pela decomposição gráfica da série histórica em suas componentes de tendência, sazonalidade e aleatória, a fim de obter uma interpretação visual da mesma. Esta decomposição é feita por meio de filtros de médias móveis (Barros et al, 2018). Também foi feita uma análise descritiva apresentando medidas de posição (valores mínimo e máximo, média e mediana), além de estatísticas que quantificam a dispersão dos dados (desvio padrão e coeficiente de variação).

Posteriormente, verificou-se a necessidade de transformação logarítmica, em virtude de uma possível instabilidade na variância. Esta estabilidade foi avaliada por meio de teste da hipótese de nulidade de que o

coeficiente da inclinação da reta amplitude versus média é zero. A necessidade de transformação ocorre quando o teste apresentar um valor-p inferior a 0,05. Se não houver necessidade de tal transformação, um modelo aditivo pode ser utilizado para captar a dinâmica temporal da série. Nesse caso, a sazonalidade não interfere na análise da componente tendência, permitindo-se analisar esta última primeiramente.

A estacionariedade da série foi investigada através de três testes, visto que os testes podem divergir na decisão a ser tomada. Isso se dá em virtude do quão rigorosos eles são, sendo necessário identificar o teste que melhor retrata o comportamento das características da série em análise. Os testes utilizados foram Cox Stuart (Box & Jenkins, 1976), de Run (Morettin & Toloi, 2006) e Dickey Fuller Aumentado (Dickey & Fuller, 1981).

O teste de Run investiga se uma série de dados foi gerada aleatoriamente, ou seja, não possui tendência. O teste de Cox Stuart é utilizado para avaliação de séries monótonas. Já o teste Dickey-Fuller Aumentado (DFA) avalia a presença de raiz unitária.

Em 1979, Dickey e Fuller propuseram um teste para verificar a presença de raiz unitária e detectar se a uma dada série foram tomadas tantas diferenças quantas forem suficientes para torná-la estacionária (Dickey & Fuller, 1979). O teste parte do pressuposto que o processo gerador dos dados é um autoregressivo de ordem um, AR (1). Entretanto, para algumas séries temporais os resíduos do teste Dickey-Fuller podem ser autocorrelacionados, tornando inadequada a representação do processo estocástico como um AR (1).

Assim, em 1981, Dickey e Fuller desenvolveram outro teste, conhecido por teste Dickey Fuller Aumentado (DFA), adicionando os valores defasados ao modelo. Na prática o teste é realizado para a série em primeira diferença, incorporando uma constante de nível e um termo de tendência determinística, a partir de três equações de regressão:

$$\Delta Z_t = \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Z_{t-1} + \varepsilon_t \quad (1)$$

se Z_t é um passeio aleatório;

$$\Delta Z_t = \beta_1 + \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Z_{t-1} + \varepsilon_t \quad (2)$$

se Z_t é um passeio aleatório com deslocamento e

$$\Delta Z_t = \beta_1 + \beta_{2t} + \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Z_{t-1} + \varepsilon_t \quad (3)$$

se Z_t é um passeio aleatório com deslocamento em torno de uma tendência determinística.

se Z_t é um passeio aleatório com deslocamento em torno de uma tendência determinística.

Em todas elas, εt, é um ruído branco, $Z_{t-1} = Z_{t-i} - Z_{t-i-1}$, β_1 a constante de nível, β_2 o coeficiente do termo de tendência determinística e α_io coeficiente da i-ésima defasagem incorporada ao modelo. O teste avalia se o processo apresenta uma raiz unitária por meio da hipótese de nulidade H0: $\delta = 0$. Se o valor-p obtido para o teste for inferior a 5% conclui-se que a série é estacionária. Os termos: raiz unitária, passeio aleatório, não estacionariedade e tendência estocástica são considerados sinônimos (Gujarati & Potter, 2008). Também são testadas hipóteses sobre os coeficientes β_1 e β_2, a fim de avaliar a presença da constante e da componente de tendência determinística no modelo, definindo, assim, o processo estocástico. O processo será um passeio aleatório com deslocamento em torno de uma tendência determinística se os valores-p obtidos para os testes de β_1 e β_2 forem menores que 0,05. A Tabela 1 apresenta como o processo é realizado e quais as hipóteses testadas em cada modelo. Neste trabalho foram utilizadas duas defasagens.

Tabela 1 - Resumo do teste Dickey Fuller Ampliado

Modelo	Hipótese
$\Delta Z_t = \beta_1 + \beta_2 t + \delta Z_{t-1} + \varepsilon_t$	$\delta = 0$
	$\delta = \beta_2 = 0$
	$\beta_1 = \delta = \beta_2 = 0$
$\Delta Z_t = \beta_1 + \delta Z_{t-1} + \varepsilon_t$	$\delta = 0$
	$\beta_1 = \delta = 0$
$\Delta Z_t = \delta Z_{t-1} + \varepsilon_t$	$\delta = 0$

Fonte - Adaptado de Enders (2008)

Segundo Enders (2008, apud Paiva, 2020) uma série contendo uma raiz unitária torna-se estacionária por diferenciação. Então, uma diferença foi aplicada à série do IEC.

Embora neste trabalho tenham sido utilizados os três testes mencionados para avaliar estacionariedade, existem outros testes utilizados para este fim, tais como Mann-Kendall, baseado na inclinação de Theil-Sen (Rutkowska, 2015), o teste baseado no coeficiente de correlação de Spearman (Sonali & Kumar, 2013) e o teste de Zivot Andrews (Zivot & Andrews, 1992).

Na sequência avaliou-se a existência de sazonalidade. Para identificar o período sazonal foi feita a decomposição espectral e a construção do Periodograma, o qual consiste na decomposição da série temporal em uma série de Fourier (Morettin & Toloi, 2006). Para avaliar a significância do período para o qual obteve-se a maior densidade espectral, foi realizado o teste G de Fisher, sob a hipótese nula de não existência de sazonalidade. A estatística G é obtida pela Equação 4, a qual é o quociente entre a maior densidade espectral $Ip(f_i)$, sendo *fi* a frequência, e a soma de todas as densidades espectrais. Se o valor-p obtido para o teste for menor que 5%, conclui-se que há sazonalidade de período $s=1/f_i$.

$$G = \frac{max\left[I_p(f_i)\right]}{\sum_{i=1}^{\left(\frac{n}{2}\right)} I_p(f_i)} \tag{4}$$

Assim como ocorre com a tendência, existem outras formas de avaliar a sazonalidade. Os leitores interessados na detecção de sazonalidade a partir de métodos de regressão linear múltipla com variáveis *Dummies* podem consultar mais detalhes em Barbosa et al. (2015), Morettin & Toloi (2006) e Gujarati & Porter (2011); os interessados no método da Análise de Variâncias e no teste de Kruskal-Walis podem pesquisar em Findlley et al (1998).

Ao que se refere à modelagem, os modelos utilizados para descrever séries temporais são processos estocásticos que necessitam considerar a auto correlação existente nos erros observados e integrar flutuações, que por vezes ocorrem nas séries, refletindo comportamentos não estacionários (tendências). Nesse sentido, a classe de modelos autoregressivos, integrados e de médias móveis ARIMA (Autoregressive Integrate Moving Average) é útil. Os modelos incluem parâmetros autoregressivos de ordem p, parâmetros de médias móveis de ordem q e são integrados na medida em que permitem remover da série a tendência por meio de d diferenças, sendo, por isso, representado por ARIMA (p, d, q).

Entretanto, diversas séries temporais apresentam padrões de repetição a cada intervalo de tempo. Como exemplo, pode-se citar as indústrias; tanto as

vendas quanto a produção seguem forte sazonalidade em certas épocas do ano (Walter et al., 2013). Esses são padrões de comportamento sazonal e não são contemplados na classe de modelos ARIMA. Mediante esta necessidade, Box e Jekins propuseram, em 1976, a classe de modelos SARIMA (Seasonal Autoregressive Integrate Moving Average), na qual foram incluídos parâmetros sazonais autoregresssivos de ordem P e de médias móveis de ordem Q, além das D diferenças sazonais necessárias para tornar a série estacionária. O modelo foi nomeado SARIMA (p, d, q) (P, D, Q)$_S$ e descreve a variável resposta Z_t pela seguinte equação:

$$\phi_p(B)\Phi_P(B^S)\Delta^d\Delta_S^D Z_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (5)$$

em que:

i. ε_t é o resíduo do modelo,

ii. $\Delta^d = (1-B)^d$ e $\Delta_S^D = (1-B^S)^D$ são operadores diferença, sendo *d* e *D* o número de diferenças necessárias, respectivamente, à eliminação da tendência e da sazonalidade;

iii. $\phi_p(B) = 1 - \phi_1(B) - \phi_2(B^2) - \cdots - \phi_P(B^p)$, o polinômio autoregressivo de ordem *p;*

iv. $\theta_q(B) = 1 - \theta_1(B) - \theta_2(B^2) - \cdots - \theta_q(B^q)$, o polinômio médias móveis de ordem *q*;

v. $\Phi_p(B^S) = 1 - \Phi_1(B^S) - \Phi_2(B^{2S}) - \cdots - \Phi_P(B^{PS})$, e $\Theta_Q(B^S) = 1 - \Theta_1(B^S) - \Theta_2(B^{2S}) - \cdots - \Theta_Q(B^{QS})$, são, nesta ordem, os operadores autoregressivo e de médias móveis sazonais;

A aplicação da metodologia de Box e Jenkins se desenvolve em três etapas: identificação, estimação e diagnóstico. A fase de identificação dos parâmetros ocorre através das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), observando as defasagens (lags) significativas estatisticamente. Dessa forma obtém-se a ordem do modelo. Porém, devido ao grande número de modelos possíveis de serem identificados, este pode ser um processo complexo e subjetivo. Neste trabalho considerou-se modelos ARIMA e SARIMA de várias ordens.

Na segunda etapa, para cada modelo identificado, foi feita a estimação dos parâmetros pelo método da máxima verossimilhança exata, satisfazendo as condições de invertibilidade e unicidade (Rezende et al, 2005). Utilizou-se o Critério de Akaike (AIC) (Akaike, 1974) para selecionar os quatro modelos mais

parcimoniosos. Tal critério considera o número de parâmetros do modelo ajustado e sua variância a partir da seguinte equação:

$$AIC(l,k) = ln\left(\sigma_{k,l}^2\right) + 2\frac{(k+l)}{N} \quad (6)$$

em que $\sigma_{k,l}^2$ é uma estimativa da variância residual e N o número de observações.

O diagnóstico dos modelos foi verificado pela análise dos resíduos. Estes foram avaliados quanto à autocorrelação serial por meio do teste estatístico Ljung-Box (Ljung & Box, 1978) e da função de autocorrelação, sendo considerado adequado o modelo que apresentou no máximo 5% de lags estatisticamente significativos.

Por fim, foram feitas previsões fora da amostra, para o horizonte de 6 meses, isto é, para o período de agosto de 2020 a janeiro de 2021. A previsão de observações temporais é a estimativa da verossimilhança de eventos futuros, baseados na informação atual e histórica (Ehlers, 2007; Souza & Camargo, 2004). A fim de obter o erro percentual absoluto médio (MAPE), foram realizadas previsões dentro da amostra, para o período de agosto de 2019 a julho de 2020, visando captar a queda ocorrida neste período devido à

pandemia. O MAPE é obtido por meio da equação:

$$MAPE = \frac{\sum_{i=1}^{k}\left[\frac{(Z_t - F_t)}{Z_t}\right]}{N} \quad (7)$$

na qual Z_t é o valor real observado, F_t o valor previsto no período t e N o número de períodos de previsão t.

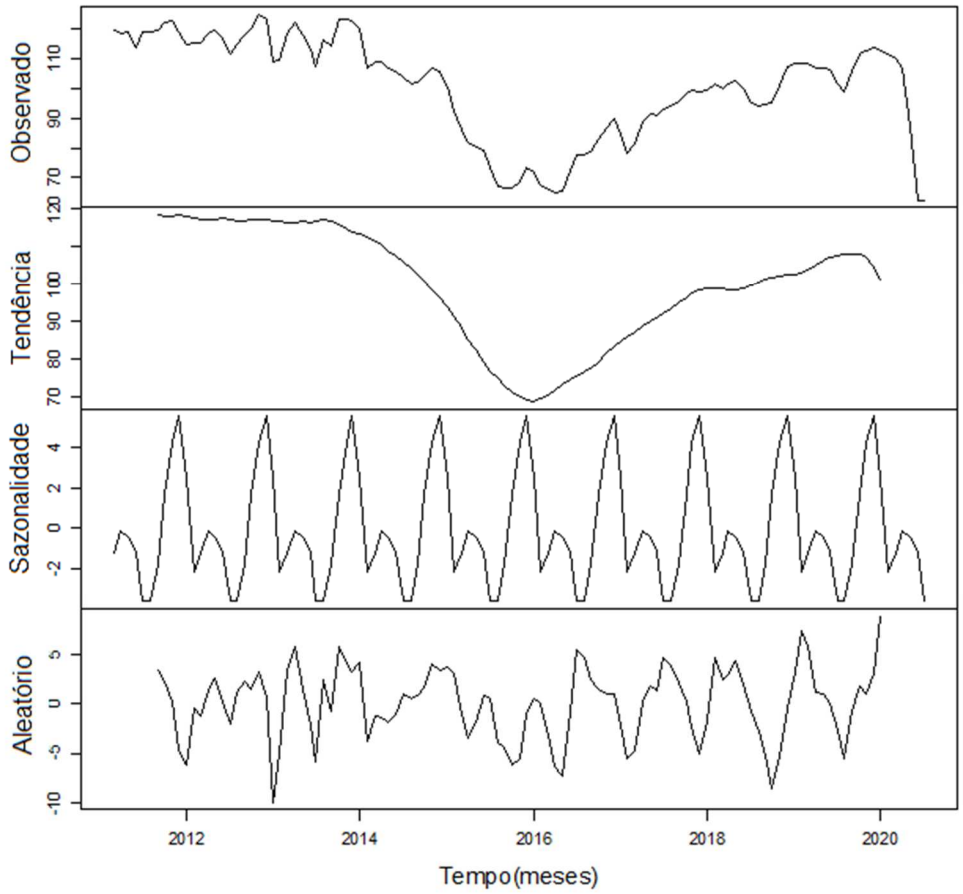
A análise estatística foi realizada com a utilização dos softwares livres GRETL (Cottrell, 2020) e R (2020), versão 4.0.2. No R foram utilizados principalmente os pacotes tseries, urca e forecast.

3. RESULTADOS

A Figura 1 apresenta o gráfico da série do índice de expansão do comércio no período de março de 2011 a julho de 2020 e sua decomposição nas componentes tendência, sazonalidade e aleatória. Observa-se nessa figura um padrão periódico bem definido subentendendo a existência de sazonalidade. Nota-se que a série não é estacionária; apresenta um decrescimento a partir de dezembro de 2013 até o ano de 2016, quando então retoma o crescimento, pressupondo a presença de tendência estocástica. Mais ainda, a partir de dezembro de 2019 o índice apresenta uma queda, o que pode ser relacionado

ao surgimento do vírus SARS-CoV-2, o qual deu início, na China, à pandemia da COVID-19, atingindo diversos países, inclusive o Brasil.

Figura 1 - Gráfico da série temporal do índice de expansão do comércio e sua decomposição



Fonte – Elaboração Própria

Para a análise da necessidade de transformação, o teste t realizado apresentou um valor p de 0,092, indicando assim que o modelo é aditivo. Desse modo, uma transformação do tipo logarítmica não foi necessária.

As medidas descritivas obtidas para a série são apresentadas na Tabela 2. O desvio padrão e o coeficiente de variação mostram a variabilidade do índice de expansão do comércio em relação ao índice médio de 100,479. Os valores não são muito baixos e corroboram para a suspeita de existência de uma tendência da série. Relativamente à mediana, observa-se uma proximidade entre o seu valor e o valor da média sinalizando na direção de uma leve assimetria dos dados. Durante os 9 anos no qual o índice foi analisado, seu menor valor foi de 62,5 (Julho de 2020) e o maior de 124,8 (Novembro de 2012).

Tabela 2 - Estatísticas descritivas para a série do índice de expansão do comércio

Medidas	Mínimo	Máximo	Mediana	Média	Desvio Padrão	Coeficiente de Variação
Valor	62,500	124,800	105,300	100,479	17,181	17,099%

Fonte – Elaboração Própria

Os resultados para o Teste de Dickey-Fuller Aumentado (ADF), incluindo duas defasagens, encontram-se na Tabela 3. Verifica-se que o coeficiente de tendência determinística (valor-p = 0,3118 para β_2) não foi significativo, indicando um modelo sem essa componente, tal como ocorre com a constante (valor-p = 0,1727 para β_1). Então, considerando Z_t um passeio aleatório, observa-se que a série não é estacionária em nível, pois o teste DFA não rejeitou a hipótese nula de raiz unitária nos dados (valor-p = 0,2916 para δ). Foi necessária apenas uma defasagem para que os resíduos do teste se tornassem um ruído branco.

Os testes de Cox Stuart e de Run apresentaram ambos valor p inferior a 0,001. Dessa forma, os três testes apontam para a não estacionariedade da série e a necessidade de integrar a série.

Tabela 3 – Resultados para o teste Dickey Fuller Aumentado

Modelos	Parâmetro	Estimativa	Erro Padrão	Valor p
Passeio aleatório com deslocamento em torno de uma tendência	β_1	5,88207	3,44327	0,0905
	β_2	-0,01528	0,01504	0,3118
	δ	0,46058	0,09644	< 0,0001 ***
	α_1	0,46058	0,09644	< 0,0001 ***
	α_2	-0,21232	0,11061	0,0576
Passeio aleatório com deslocamento	β_1	3,78382	2,75600	0,1727
	δ	-0,04110	0,02696	0,1304
	α_1	0,46347	0,09641	< 0,0001 ***
	α_2	-0,22661	0,10973	0,0413 *
Passeio aleatório	δ	-0,00456	0,004299	0,2916
	α_1	0,445028	0,095867	< 0,0001 ***
	α_2	-0,25321	0,10845	0,0214 *

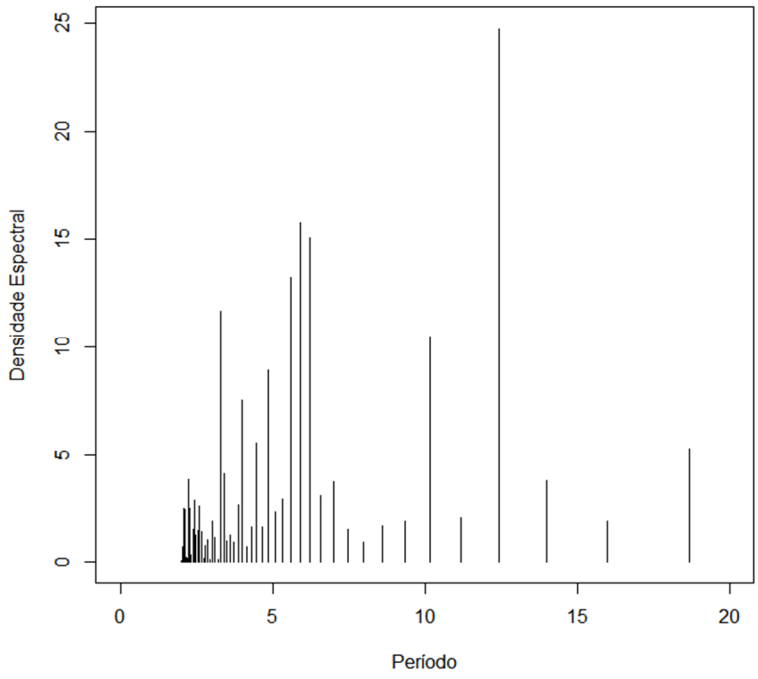
Fonte – Elaboração Própria

Significativo a: * 5% de probabilidade, ***1% de probabilidade

De acordo com a Tabela 3, para o modelo passeio aleatório as duas primeiras defasagens são significativas. Porém, para a série com uma defasagem, a estimativa da autocorrelação de ordem 1 é 0,3530 e para a série com duas defasagens, essa estimativa é de -0,19. O valor negativo dessa estimativa sinaliza, segundo Morettin & Toloi (2006), um excesso de defasagens. Além disso, o teste DFA e o teste de Cox Stuart aplicados à série defasada de ordem um sinalizaram que essa série é estacionária. Então, a ordem de integração da série foi um.

Em virtude da suspeita de sazonalidade nos dados vinda da observação do padrão periódico existente, realizou-se a decomposição espectral, apresentada na Figura 2, a fim de identificar o período de ocorrência. Observou-se no Periodograma que a maior densidade espectral corresponde ao período de 12 meses. O Teste G de Fisher apresentou um valor-p de 0,076, não identificando, ao nível de 5% de significância, a presença de uma sazonalidade de 12 meses.

Figura 2 – Periodograma da série IEC integrada de ordem um

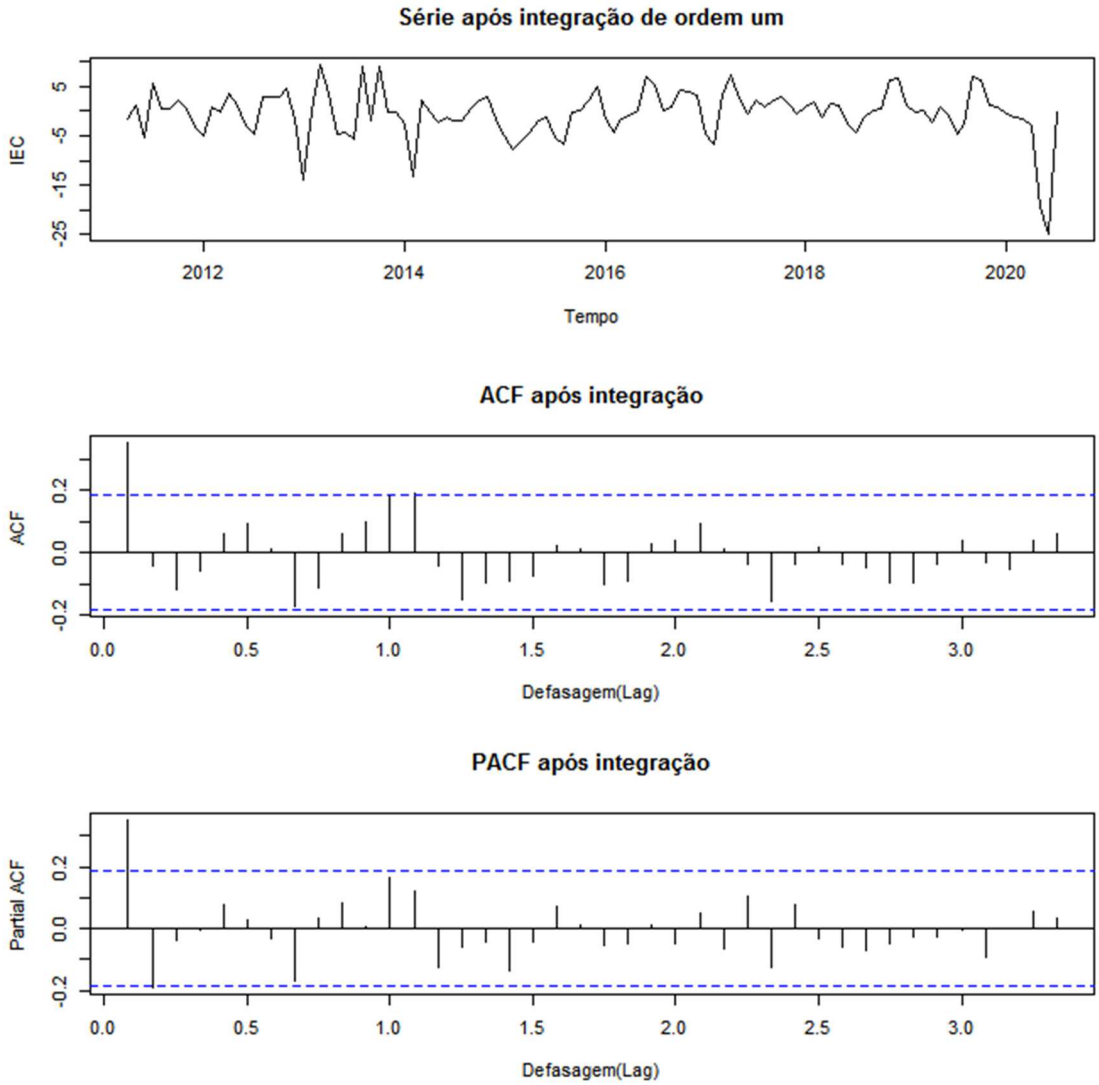


Fonte – Elaboração Própria

A Figura 3 ilustra a série com uma diferença de ordem um e suas respectivas funções de autocorrelação (FAC) e autocorrelação parcial (PACF). Observa-se que o lag 12 na FAC é significativo, o que, de acordo com Morettin & Toloi (2006), aponta para uma sazonalidade estocástica, a qual deve ser eliminada através do ajuste de um modelo que a considere, no caso um modelo

da classe SARIMA $(p,d,p)(P,D,Q)_{12}$. Assim, diversos modelos foram ajustados, considerando d = 1, em virtude da não estacionariedade da série em nível e D = 0, pois o teste G de Fisher não identificou sazonalidade determinística. Uma vez que na FAC o lag 12 é significativo e na FACP não, os modelos foram tomados apontando os valores 0 e 1 para Q e 0 para P. Além disso, os lags 1 na FAC e 1 e 2 na FACP também são significativos, indicando que os valores de p devem variar entre 0 e 1 e os valores de q devem variar de 0 a 2.

Figura 3 – Série temporal, ACF e PACF após aplicação da diferença



Fonte – Elaboração Própria

Os modelos cujos parâmetros apresentaram coeficientes significativos de acordo com o teste t de Student estão listados na Tabela 4. Para todos eles os resíduos são ruído branco, visto que possuem valor-p superior a 5% para o teste Ljung-Box. Os modelos ARIMA (0,1,1) e ARIMA (2,1,0) apresentaram os menores valores do critério AIC. Porém, foi para o modelo ARIMA (2,1,0) que se

obteve o menor MAPE, sendo este então o escolhido na modelagem dos dados do IEC.

Tabela 4 - Comparação entre os modelos ajustados

Modelo	Ljung-Box (valor-p)	MAPE	AIC
SARIMA (0, 1, 1) x $(0, 0, 1)_{12}$	0,294	7,0111	674,0658
ARIMA (0,1,1)	0,770	6,2799	661,2539
ARIMA (1,1,0)	0,454	6,7481	664,8552
ARIMA (2,1,0)	0,853	6,0999	661,4699

Fonte – Elaboração Própria

A Tabela 5 mostra as estimativas dos coeficientes dos parâmetros e suas estatísticas.

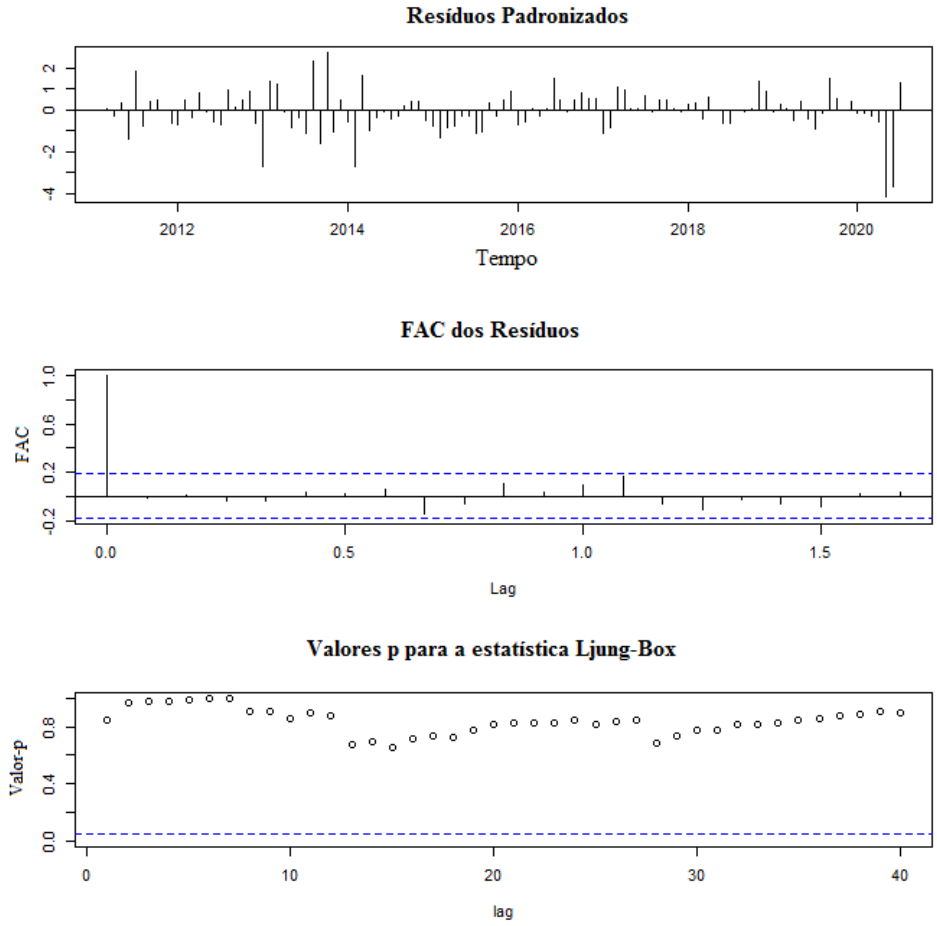
Tabela 5 - Estatísticas para o modelo selecionado ARIMA (2,1,0)

Parâmetro	Coeficiente	Erro Padrão	Valor p
f_1	0,446251	0,0937667	< 0,0001
f_2	-0,247914	0,105367	0,0186

Fonte – Elaboração Própria

Para avaliação do modelo selecionado, foram construídos gráficos para os resíduos, os quais se encontram na Figura 4. De acordo com os resíduos padronizados, os dados se encontram distribuídos simetricamente em torno de média zero. A função de autocorrelação (ACF) não indica presença de autocorrelação nos resíduos, visto que não há lags significativos. Por fim, o gráfico dos valores-p para a estatística Ljung-Box até a defasagem 40 sinaliza que os resíduos são ruído branco, pois estes valores estão acima de 0,5.

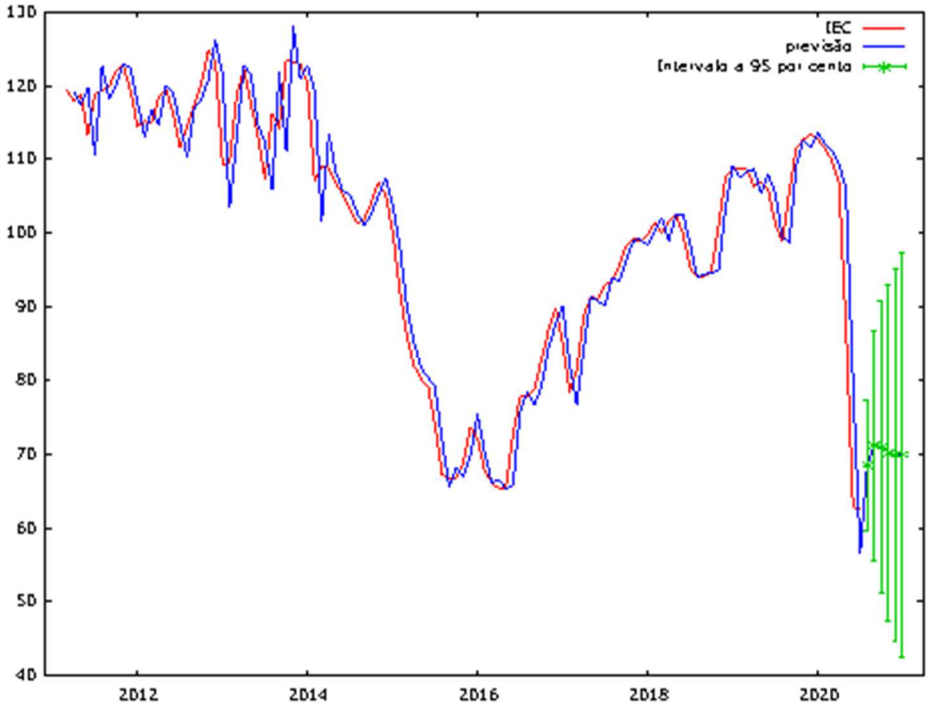
Figura 4 - Gráficos para avaliação dos resíduos



Fonte – Elaboração Própria

A Figura 5 ilustra o ajuste do modelo ARIMA (2,1,0) aos dados do IEC e sua previsão para 12 meses, bem como a série real. Nota-se uma proximidade entre as curvas, indicando que o modelo ajustado captou o comportamento temporal da série.

Figura 5 - Gráfico para a série real do IEC (vermelho), o modelo ajustado (azul) e suas previsões para seis meses com respectivos intervalos de 95% de confiança (verde)



Fonte – Elaboração Própria

Na Tabela 6 são apresentadas algumas informações referentes às previsões para os meses de agosto de 2019 a julho de 2020.

Tabela 6 – Estimativas preditas pelo modelo ARIMA (2,1,0) e respectivos erros de previsão

Mês/Ano	Valor Previsto	Valor real	Erros de Previsão
Agosto /2019	99,6	98,8	-0,8
Setembro/2019	98,8	105,6	6,8
Outubro/2019	109,3	111,7	2,4
Novembro/2019	112,7	112,7	0,0
Dezembro/2019	111,6	113,5	1,9
Janeiro/2020	113,6	112,8	-0,8
Fevereiro/2020	112,3	111,6	-0,7
Março/2020	111,2	109,9	-1,3
Abril/2020	109,4	107	-2,4
Maio/2020	106,1	87,5	-18,6

Junho/2020	79,5	62,8	-16,7
Julho/2020	56,6	62,5	5,9

Fonte – Elaboração Própria

Para finalizar e avaliar a qualidade das previsões, foi calculado o erro percentual absoluto médio de previsão (MAPE), para o qual obteve-se o valor de 6,099%. O baixo valor obtido (inferior a 15%) indica que as previsões são boas. Percebe-se que as previsões são de elevação do IEC. Observa-se que em julho de 2020 houve redução da taxa Selic, a qual atingiu seu menor valor nos últimos anos. A redução dessa taxa gera juros mais baixos no mercado, incentivando os investimentos empresariais. Além disso, as medidas de isolamento social estão sendo flexibilizadas, incentivando o consumo.

4. CONSIDERAÇÕES FINAIS

A análise do índice de expansão do comércio no período de junho de 2011 a julho de 2020 revelou que a série apresenta uma tendência estocástica e não possui sazonalidade determinística. Essas características indicaram a necessidade de se utilizar modelos integrados capazes de captar o comportamento da série e estimar valores futuros que possam auxiliar decisões governamentais e empresariais.

Foi utilizada a metodologia proposta por Box & Jenkins (1976) e os modelos cujos resíduos são ruído branco foram selecionados segundo o critério de Akaike, obtendo o modelo mais parcimonioso e garantindo a adequacidade do ajuste. Dentre os quatro modelos selecionados, utilizou-se o modelo ARIMA (2,1,0) para captar a dinâmica temporal da série e realizar previsões, em virtude de seu menor erro percentual absoluto médio. Os valores estimados no período de agosto de 2019 a julho de 2020 foram próximos aos valores reais, fornecendo um MAPE de aproximadamente 6%. De acordo com o modelo ARIMA (2,1,0) as previsões para os próximos meses apontam um crescimento do IEC, embora modestos se comparados aos últimos três anos.

A metodologia empregada mostrou-se adequada para a modelagem do IEC. Assim, pretende-se, em trabalhos futuros, utilizar os modelos propostos por Box e Jenkins no estudo desse índice, incluindo análise de intervenções nos períodos próximos a dezembro de 2013, maio de 2016 e principalmente dezembro de 2019. Isto porque os reflexos da pandemia da COVID 19, ocorrida em 2019/2020 podem permanecer por tempo ainda indeterminado. Além disso, sugere-se a aplicação dessa metodologia em dados de IEC em outros estados brasileiros, a fim de que os poderes públicos estaduais, bem como a classe empresária, possam se beneficiar dos resultados a serem obtidos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Akaike, H. (1974). A new look at the statistical model identification. IEE Transactions on Automatic Control, 19, 716-723. . doi: 10.1109/TAC.1974.1100705

Aquino, E. M., Silveira, I. H., Pescarini, J. M., Aquino, R., Souza-Filho, J. A. D., Rocha, A. D. S. & Lima, R. T. D. R. S. (2020). Medidas de distanciamento social no controle da pandemia de COVID-19: potenciais impactos e desafios no Brasil. Ciência & Saúde Coletiva, *25*, 2423-2446.

Barbosa, E. C., Sáfadil, T., Nascimento, M., Nascimento, A. C. C., Silva, C. H. O & Manuli, R. C. (2015). Metodologia Box & Jenkins Para Previsão De Temperatura Média Mensal Da Cidade De Bauru (Sp). Revista Brasileira de Biometria, 33(1), 104–117.

Barros, A. C., Mattos, D. M., Oliveira, I. C. L., Ferreira, P. G. C. & Duca, V. E. L. A. (2018). Análise de séries temporais em R: curso introdutório. (1a. ed.). Rio de Janeiro: Elsevier FGV IBRE.

Box, G. & Jenkins, G. (1976). Time series analysis, forecasting and control. San Francisco: Holden-Day.

Cottrell, A. (2020). Gretl - Gnu Regression, Econometrics and Time-series library. Wake Forest University. Disponível em: < http://gretl.sourceforge.net/win32/index_pt.html >

Dickey, D. A. & Fuller, W. A. (1979). Distributions of the estimators for autoregressive. time series with a unit root. Journal of the American Statistical Association, 75, 427–431. doi: 10.2307/2286348.

Dickey, D. A. & Fuller, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. Econometrica, 49, 1057–1072. doi: 10.2307/1912517

Ehlers, R. S. (2007). Análise de séries temporais. Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná. Disponível em <http://www.each.usp.br/rvicente/AnaliseDeSeriesTemporais.pdf> Acesso em 21/10/2021.

Enders, W. (2008). Applied econometric time series. (4a ed). EUA: John Wiley & Sons.

FECOMERCIOSP. Índice de Expansão do Comércio. Disponível em: < http://www.fecomercio.com.br/pesquisas/indice/iec>. Acesso em: 26/08/2020.

_____. Relatório Anual 2018. Disponível em: https://www.fecomercio.com.br/upload/file/2019/06/24/raf2018.pdf. Acesso em: 07/08/2020.

Fernandes, F. (2021). Empresários de SP estão mais otimistas com os negócios. Disponível em: <https://dcomercio.com.br/categoria/economia/empresarios-de-sp-estao-mais-otimistas-com-os-negocios> Acesso em 21/10/2021.

Findley, D. F., Monsell, B.C., Bell, W.R., Otto, M.C. & Chen, BC. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. Journal of Business & Economic Statistics, 16 (2), 127-152. doi: 10.2307/1392565

Gujarati, D. N. & Porter, D. C. (2011). Econometria Básica. (5a. ed). Porto Alegre: Amgh Editora.

INVESTSP. Agência Paulista de Promoção de Investimentos e Competitividade. Comércio em São Paulo. s/d. Disponível em: < https://www.investe.sp.gov.br/por-que-sp/economia-diversificada/comercio/>. Acesso em 24/08/2020.

INVESTSP. Agência Paulista de Promoção de Investimentos e Competitividade. PIB. Disponível em: < https://www.investe.sp.gov.br/por-que-sp/economia-diversificada/pib/>. Acesso em 21/10/2021.

Ljung, G. M. & Box, G. E. (1978). On a measure of lack of fit in time series models. Biometrika, 65 (2), 297-303. doi: 10.2307/2335207

Morettin, P. A. & Toloi, C. M. C. (2006). Análise de séries temporais (2a .ed). São Paulo: Edgard Blucher.

Oikawa, K. F. (2020). Os efeitos da crise financeira americana de 2008 na produção física industrial brasileira: uma análise de intervenção. Revista Brasileira de Estatística, 78 (243), 51-80.

Oliveira, M. L. M. (2020). Confiança: a base da economia em 2020. Disponível em < https://diariodocomercio.com.br/opiniao/confianca-a-base-da-economia-em-2020/> Data do acesso: 21 de outubro de 2021.

Paiva, D. A. (2020). Estudos de testes para tendência em séries temporais. (Dissertação de mestrado). Universidade Federal de Lavras (UFLA), Lavras.

R Development Core Team (2021). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rezende, J. L. P., Coelho Junior, L. M., Oliveira, A. D. & Sáfadi, T. (2005). Análise dos preços de carvão vegetal em quatro regiões no estado de Minas Gerais. CERNE, 11, 237-252.

Rodrigues, F. PIB da cidade de São Paulo equivale ao de 4.305 outros municípios. 2019. (2019). Disponível em <https://www.poder360.com.br/brasil/pib-da-cidade-de-sao-paulo-equivale-ao-da-soma-de-4-305-municipios/> Acesso em 21/10/2021.

Rutkowska, A. (2015). Properties of the Cox-Stuart test for trend in application to hydrological series: the simulation study. Communications in Statistics: Simulation and Computation, 44 (3), 565–579. https://doi.org/10.1080/03610918.2013.784988

Silva, W. P. C., Nascimento, F. F. & Ferraz, V. R. S. (2020). Uso de ferramentas econométricas para estimar e modelar o PIb do Brasil. Revista Brasileira de Estatística, 78 (243), 81-111.

Sonali, P. & Kumar, D. N. (2013). Review of trend detection methods and their application to detect temperature changes in India. Journal of Hydrology, 476, 212–227. Disponível em: < https://doi.org/10.1016/j.jhydrol.2012.10.034>

Souza, R. C. & Camargo, M. E. (2004). Análise e previsão de séries temporais: os modelos ARIMA. (2a. ed). Rio de Janeiro: Regional.

Walter, O. M. F.C., Henning, E., Moro, G. & Samohyl, R. W. (2013) Aplicação de um modelo SARIMA na previsão de vendas de motocicletas. Exacta, 11 (n. 1), 77–88, doi: 10.5585/ExactaEP.v11n1.3897

Zivot, E. & Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. Journal of Business & Economic Statistics, 10 (3), 25–44. doi: 10.2307/1391541.

Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. J. (2020). Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology, *296* (2), E15-E25.

Revista Brasileira de Estatística, V.80, N. 247, p.48-74, 2022

CONDICIONANTES DO CENSO AGROPECUÁRIO 2017

Vicente Penteado Meirelles de Azevedo Marques

vicente.marques@incra.gov.br

Instituto Nacional de Colonização e Reforma Agrária (INCRA)

Resumo: Este trabalho tem como objetivo apresentar e discutir elementos metodológicos para subsidiar as análises comparativas entre os resultados do Censos Agropecuários de 2006 e 2017 no Brasil, especialmente em relação aos estabelecimentos da Agricultura Familiar. Os resultados obtidos sugerem possibilidades de associação da diminuição do número desses estabelecimentos e os condicionantes analisados.
Palavras-chave: metodologia; estatística; agricultura familiar

Abstract: This paper aims to present and discuss methodological elements to support comparative analyses between the results of the 2006 and 2017 Agricultural Censuses in Brazil, especially in relation to Family Farming establishments. The results obtained suggest possibilities of association of the decrease in the number of these establishments and the conditioning factors analysed.
Keywords: methodology; statistics; family farming

1. INTRODUÇÃO

Um dos princípios fundamentais do Censo Agropecuário (CA) é a busca pela manutenção de uma coerência das suas características com a intenção de seus resultados serem comparáveis ao longo do tempo. Dessa preocupação decorrem os constantes alertas do IBGE a respeito dos cuidados na abordagem metodológica a serem adotados para a correta interpretação dos dados de diferentes pesquisas. Apesar disso, as implicações destes procedimentos são pouco conhecidas.

Entre essas implicações podem ser citados os efeitos das mudanças nos períodos de coleta e de referência dos censos sobre a não-captação de estabelecimentos de natureza precária e com atividade temporária, como aqueles trabalhados por arrendatários(as) e parceiros(as) em pequenas áreas ou então, por proprietários(as) que residem longe do estabelecimento e comparecem a ele ocasionalmente na entressafra, sem possibilidade de identificação pela pesquisa (Hoffmann, 2007).

Este texto tem como objetivo fornecer elementos para subsidiar as análises comparativas entre os resultados dos CA de 2006 e 2017, especialmente em relação aos estabelecimentos familiares. Nesse período existem evidências de um número significativo de unidades produtivas que deixou a categoria legal Agricultura Familiar, sendo que muitos destas migraram para a categoria não familiar, especialmente os(as) produtores(as) com menor renda monetária (Del Grossi et al., 2019).

A primeira parte do texto descreve as principais mudanças no método do CA 2017 em relação à pesquisa anterior. A segunda parte caracteriza alguns elementos que podem estar vinculados aos possíveis impactos dessas mudanças observadas e apresenta dados comparativos quantitativos entre as duas pesquisas. A terceira parte comenta outros condicionantes associados aos elementos mencionados na parte anterior e a quarta parte tece considerações finais e apresenta novas hipóteses para melhor compreensão dos resultados do CA 2017.

2. PRINCIPAIS MUDANÇAS DO CENSO AGROPECUÁRIO 2017

As principais mudanças do CA 2017 em relação ao CA 2006 foram as seguintes:

a) Período de coleta: de maio a setembro de 2007 para 1º de outubro de 2017 a 28 de fevereiro de 2018. O período de coleta corresponde ao intervalo em que os dados censitários foram obtidos junto às pessoas entrevistadas. As

datas indicadas são aproximadas devido às diferenças verificadas em cada Unidade da Federação.

b) Data de referência: de 31 de dezembro de 2006 para 30 de setembro de 2017. A data de referência é utilizada nas variáveis total de estabelecimentos agropecuários; área total dos estabelecimentos agropecuários; utilização das terras; total de tratores, implementos, máquinas e veículos; características do estabelecimento e do produtor; total de pessoal ocupado; e totais de bovinos, bubalinos, caprinos, suínos, ovinos, equinos, muares e aves.

c) Período de referência: de 1º de janeiro a 31 de dezembro de 2006 para 1º de outubro de 2016 a 30 de setembro de 2017. O período de referência é utilizado nas variáveis referentes à produção animal (leite, lã, ovos de galinha e de outras aves etc.) e vegetal, às práticas agrícolas (utilização de agrotóxicos, adubos, irrigação etc.), número máximo de pessoas ocupadas e a valores da produção, receitas e despesas associadas aos processos produtivos.

d) Produção/criação de empregados/moradores em área do estabelecimento agropecuário. Para minimizar os efeitos da subenumeração dos estabelecimentos temporários decorrente da alteração dos períodos de referência, nos dois censos mais recentes o IBGE tem conceituado e identificado os “produtores(as) sem área”. No entanto, de forma diferente do censo anterior, o CA 2017 não abriu o questionário de novo estabelecimento para o(a) produtor(a) empregado(a)/morador(a) em área sujeita à administração do(a) produtor(a)/proprietário(a). Desta forma, toda a produção/criação referida a esta condição integrou o questionário do estabelecimento agropecuário do produtor(a)/proprietário(a). Até o CA 2006 a produção vegetal ou animal obtidas por empregados(as) ou moradores(as) foram registradas em questionário próprio sem que isso caracterizasse a existência de novo estabelecimento agropecuário ou “produtor(a) sem área”.

e) Estabelecimento agropecuário formado por área não contínua. No CA 2017 as áreas não contínuas exploradas por um mesmo produtor foram consideradas como um único estabelecimento, desde que estivessem situadas no mesmo município, utilizassem os mesmos recursos técnicos (máquinas, implementos e instrumentos agrários, animais de trabalho etc.) e os mesmos recursos humanos (mesmo pessoal), e, também, estivessem subordinadas a uma única administração: a do(a) produtor(a) ou a do(a) administrador(a). No CA 2006, bastava que as áreas não contínuas do estabelecimento estivessem situadas em setores censitários diferentes para que fossem admitidas como

estabelecimentos distintos, consideradas algumas particularidades quanto à existência e localização de sua sede.

f) Definição de Agricultura Familiar (Lei nº 11.326/2006). Nos dois censos mais recentes a definição da Agricultura Familiar foi feita considerando os dispositivos da Lei nº 11.326/2006 e as variáveis disponíveis nos respectivos questionários. No período entre os censos essa lei foi regulamentada pelo Decreto nº 9.064/2017 e as mudanças dele decorrentes foram incorporadas à definição dos algoritmos para do CA 2017 (Del Grossi, 2019).

g) Coleta dos dados. Nos dois censos mais recentes as informações obtidas junto às pessoas entrevistadas nos estabelecimentos agropecuários foram registradas em questionários eletrônicos instalados em dispositivos móveis e transmitidas para o Banco de Dados do IBGE. Isso permitiu a introdução de rotinas de críticas de alguns dados e inserção de saltos automáticos de quadros e questões, o que contribuiu para uma melhor qualidade na coleta das informações, inclusive quanto ao maior detalhamento de algumas variáveis. A coleta do CA 2017 inovou ao ser realizada a partir de uma lista prévia de endereços com coordenadas georreferenciadas captados no CA 2006 e no Censo Demográfico 2010 e ao registrar automaticamente a posição e o percurso dos(as) recenseadores(as). Isso foi feito com o objetivo de atualizar os dados cadastrais e aperfeiçoar a supervisão das atividades de coleta e assim, conferir maior qualidade à cobertura dos estabelecimentos.

Para o dimensionamento dos possíveis impactos das mudanças metodológicas entre os dois censos mais recentes foram adotados os seguintes procedimentos:

a) Análise das características de sete lavouras temporárias (arroz, feijão preto, feijão de cor, feijão fradinho, milho, soja e trigo) quanto ao seu período de safra e a sua relação com as datas de referência e os períodos de referência e de coleta dos dois censos agropecuários mais recentes, por Unidade da Federação. Além disso, é caracterizada a condição climática, especialmente a precipitação, de cada um desses intervalos de tempo. A escolha das lavouras é justificada pelo fato de elas serem responsáveis por aproximadamente 80% do total da área colhida de lavouras temporárias, segundo a Pesquisa Agrícola Municipal (PAM), do IBGE, de 2006 e de 2017 (Tabela 1612 do Sistema IBGE de Recuperação Automática (Sidra). A escolha da precipitação é justificada pelo fato de ela ser um dos principais limitantes climáticos para o desenvolvimento de lavouras temporárias em climas tropical, seco (semiárido) e temperado (subtropical). Esta escolha tem como principal limitação a exclusão das lavouras

de mandioca e de cana-de-açúcar, cujos ciclos produtivos são maiores do que o período de referência dos censos (doze meses) e não podem ser devidamente relacionados aos parâmetros propostos. O mesmo ocorre com as lavouras permanentes, a silvicultura, a aquicultura e as atividades extrativistas.

A análise do comportamento dos efetivos e da produção da pecuária não foi realizada uma vez que os principais determinantes do número de animais no estabelecimento durante o ano estão definidos por uma grande variedade de condições de produtivas, de canais de comercialização e de ofertas de preço, o que impede uma razoável agregação dos dados por Unidade da Federação (Matte, 2017);

b) Extração de dados do Sidra e utilização de tabelas especiais disponibilizadas pelo IBGE em seu portal eletrônico. Sobre a definição dos períodos de safra das lavouras temporárias foi utilizada como referência o Calendário de Plantio e Colheita de Grãos do Brasil, da Companhia Nacional de Abastecimento (Conab) (2015), que observa as épocas recomendadas pelo Zoneamento Agrícola de Risco Climático para o plantio. Sobre os dados climáticos e sobre a colheita dos diferentes tipos de feijão foram consultados os Boletins de Grãos, também da Conab. Para os dados climáticos complementares foram consultadas as normais climatológicas (1981-2010) e os desvios da precipitação com relação a elas (anomalias) do Banco de Dados Meteorológicos (BMDEP), do Instituto Nacional de Meteorologia (Inmet), entre outras fontes. Por se tratarem de aproximações iniciais, não foram utilizados instrumentos estatísticos;

c) Os possíveis impactos das mudanças quanto aos estabelecimentos formados por área não contínua deixaram de ser analisados por serem considerados pouco relevantes para as unidades familiares, cujas parcelas invariavelmente são confinantes;

d) Para permitir a comparabilidade dos dados referentes à Agricultura Familiar foram utilizadas as variáveis do CA 2006 redefinidas de acordo a legislação vigente e as variáveis disponíveis no Censo 2017 (IBGE, 2019; Del Grossi, 2019). Estes dados são disponibilizados pelo IBGE como tabelas especiais do Censo 2006 e não constam no Sidra (IBGE, 2017b). As tabelas disponíveis têm como principal limitação o reduzido número de variáveis disponíveis.

Antes de procurar dimensionar os possíveis impactos quantitativos de cada mudança metodológica do CA, é importante identificar a referência sobre a qual será feita a maioria das considerações deste texto, ou seja, as diferenças

do número de estabelecimentos agropecuários nas duas pesquisas censitárias mais recentes.

Tabela 1- Número total de estabelecimentos agropecuários, de estabelecimentos sem área, estabelecimentos familiares (AF) e estabelecimentos familiares sem área. Brasil e Unidades da Federação. 2006 e 2017. Em unidades.

UF	2017				2006			
	Total	Sem Área	AF Total	AF Sem Área	Total	Sem Área	AF Total	AF Sem Área
Brasil	5.073.324	77.037	3.897.408	54.394	5.175.636	255.019	4.305.105	236.070
Rondônia	91.438	306	74.329	262	87.078	914	74.457	815
Acre	37.356	271	31.109	243	29.483	1.875	25.219	1.734
Amazonas	80.959	3.119	70.358	2.756	66.784	10.449	61.378	10.188
Roraima	16.846	218	13.103	193	10.310	445	8.704	442
Pará	281.699	4.126	239.737	3.809	222.029	16.093	195.595	15.559
Amapá	8.507	191	6.984	178	3.527	439	2.902	401
Tocantins	63.808	769	44.955	489	56.567	941	42.658	878
Maranhão	219.765	17.489	187.118	14.560	287.039	58.984	263.076	57.819
Piauí	245.601	8.329	197.246	5.836	245.378	24.078	218.209	22.653
Ceará	394.330	23.613	297.862	13.346	381.017	39.535	338.103	36.210
R. Grande Norte	63.452	559	50.680	418	83.053	4.379	70.412	4.029
Paraíba	163.218	1.562	125.489	1.026	167.286	7.234	145.312	6.292
Pernambuco	281.688	2.251	232.611	1.760	304.790	19.745	268.770	17.775
Alagoas	98.542	756	82.369	565	123.332	5.540	109.593	4.993
Sergipe	93.275	533	72.060	338	100.607	2.246	88.806	2.070
Bahia	762.848	6.026	593.411	3.975	761.558	19.363	647.963	17.228
Minas Gerais	607.557	2.253	441.829	1.659	551.621	14.835	432.612	12.916
Espírito Santo	108.014	628	80.775	266	84.361	598	67.534	522
Rio de Janeiro	65.224	133	43.786	77	58.493	1.912	43.500	1.807
São Paulo	188.620	772	122.555	534	227.622	2.172	148.585	1.917
Paraná	305.154	933	228.888	666	371.063	8.832	298.726	7.393
Santa Catarina	183.066	577	142.987	396	193.668	4.122	165.581	3.662
R. Grande do Sul	365.094	901	293.892	612	441.472	6.857	375.000	5.767
Mato Grosso Sul	71.164	202	43.223	135	64.864	300	40.254	236
Mato Grosso	118.679	246	81.635	146	112.987	1.016	83.031	940
Goiás	152.174	268	95.684	143	135.692	2.113	87.292	1.823
Distrito Federal	5.246	6	2.733	6	3.955	2	1.833	1

Fonte - IBGE- Censo Agropecuário (tabela Especial 1 e tabela 6878 do Sidra. Elaboração própria

Tabela 2 - Diferença do número total de estabelecimentos agropecuários e estabelecimentos familiares (AF) entre 2017 e 2006. Brasil e Unidades da Federação. Em unidades.

UF	Total Estab.	Estab. AF
Brasil	-102.312	-407.697
Rio Grande do Sul	-76.378	-81.108
Maranhão	-67.274	-75.958
Paraná	-65.909	-69.838
Bahia	1.290	-54.552
Ceará	13.313	-40.241
Pernambuco	-23.102	-36.159
Alagoas	-24.790	-27.224
São Paulo	-39.002	-26.030
Santa Catarina	-10.602	-22.594
Piauí	223	-20.963
Paraíba	-4.068	-19.823
Rio Grande do Norte	-19.601	-19.732
Sergipe	-7.332	-16.746
Mato Grosso	5.692	-1.396
Rondônia	4.360	-128
Rio de Janeiro	6.731	286
Distrito Federal	1.291	900
Tocantins	7.241	2.297
Mato Grosso do Sul	6.300	2.969
Amapá	4.980	4.082
Roraima	6.536	4.399
Acre	7.873	5.890
Goiás	16.482	8.392
Amazonas	14.175	8.980
Minas Gerais	55.936	9.217
Espírito Santo	23.653	13.241
Pará	59.670	44.142

Fonte - IBGE- Censo Agropecuário (tabela especial 1 e tabela 6878 do Sidra). Elaboração própria

A partir da Tabela 2 é possível observar que a redução do número de estabelecimentos familiares é relevante (- 9,5%) e corresponde à maior parte da diminuição do número total de unidades produtivas no País entre 2006 e 2017 (- 102 mil). Isso é especialmente perceptível na Região Nordeste (- 311 mil estabelecimentos familiares, ou – 14,5%) e na Região Sul (- 174 mil, ou – 20,7%).

No sentido oposto, houve aumento de estabelecimentos familiares na Região Norte (+ 70 mil, ou + 17,0%) e Região Centro-Oeste (+ 11 mil, ou + 5,1%). Na Região Norte as unidades familiares foram responsáveis pela maior parte do crescimento do número total de estabelecimentos (+ 105 mil). Na Região Sudeste o número de unidades produtivas familiares manteve-se praticamente estável (- 3 mil ou – 0,5%). Os estabelecimentos "sem área" são analisados na seção 2.d deste texto.

Tabela 3 - Diferença do número total de estabelecimentos agropecuários familiares (AF) entre 2017 e 2006, por grupo de atividade econômica selecionado. Brasil e Unidade da Federação. Em 1.000 unidades.

UF	Produção de lavouras temporárias	Horticultura e floricultura	Produção de lavouras permanentes	Produção de florestas plantadas	Pecuária e criação de outros animais
Brasil	-333,9	-37,3	-19,8	-19,3	12,9
Bahia	-61,9	-0,4	-0,8	-6,8	13,6
Maranhão	-49,0	-1,2	-1,5	-1,9	0,8
Ceará	-44,9	-0,5	-3,5	-2,8	6,9
Paraná	-32,5	-5,4	-6,7	-0,7	-25,3
Rio Grande do Sul	-31,5	-7,7	-0,2	-1,0	-39,1
Pernambuco	-28,8	-4,3	-2,5	-0,6	1,8
Piauí	-24,0	-1,1	-3,4	-1,0	9,6
Alagoas	-22,5	-0,7	-1,7	-0,3	-2,1
Paraíba	-19,9	-2,5	-3,0	-1,2	5,3
Santa Catarina	-19,6	-3,1	0,0	1,2	-0,1
R. Grande do Norte	-13,3	-0,1	-3,7	-0,6	-0,8
Sergipe	-9,4	0,2	-8,0	-0,4	1,0
Minas Gerais	-6,8	-9,5	7,3	-0,7	19,3
São Paulo	-6,5	0	-6,8	1,4	-13,6
Rondônia	-2,5	-0,4	-7,1	-0,3	10,6
Tocantins	-0,9	0,3	-0,3	-0,7	3,4
Mato Grosso	-0,7	-1,5	-0,2	-0,2	1,8
Rio de Janeiro	-0,2	-0,3	0	0,1	0,6
Goiás	-0,2	-0,5	-0,1	0,0	9,6
Distrito Federal	0,2	0,3	0,1	0,0	0,3
Espírito Santo	0,7	1,2	9,6	0,3	1,4
Amapá	2,0	0,4	0,2	0,0	0,3
Mato Grosso Sul	2,2	-0,4	0,0	0,0	1,4

Roraima	3,0	0,2	1,1	-0,1	-0,2
Acre	5,5	0,0	0,2	-0,5	0,7
Amazonas	11,1	-0,1	0,2	-0,4	-1,7
Pará	16,4	-0,1	11,0	-2,5	7,4

Fonte - IBGE- Censo Agropecuário (tabela especial 36 e tabela 6778 do Sidra). Elaboração própria.

A Tabela 3 revela que entre os principais grupos de Atividade Econômica pesquisados obtidos conforme a Classificação Nacional da Atividade Econômica (CNAE 2.0) (IBGE, 2019) a produção de lavouras temporárias foi aquele que apresentou maior redução do número de estabelecimentos familiares, especialmente nas Regiões Nordeste e Sul. A diminuição de unidades familiares na produção pecuária e na horticultura foi especialmente expressiva no Rio Grande do Sul e no Paraná. No sentido oposto, houve o crescimento dos estabelecimentos com pecuária e outras criações na Região Nordeste, exceto em Alagoas e no Rio Grande do Norte. Isso sugere uma possível migração da atividade da produção de lavouras temporárias para a da pecuária, como será abordado na seção 3 deste texto.

As tabelas a seguir caracterizam as lavouras selecionadas conforme as mudanças metodológicas mencionadas na primeira parte deste texto.

a) Período de coleta e coleta de dados. Durante a coleta do CA 2017 (1º de outubro de 2017 a 28 de fevereiro de 2018) encontravam-se no período entre o plantio e a colheita as seguintes lavouras:

a.1) Arroz em todas as Regiões e Estados analisados, sendo Roraima (colheita até outubro); Rio Grande do Norte (colheita até novembro) e Ceará (plantio a partir de fevereiro);

a.2) Feijões em todas as Regiões e safras – exceto Amapá (2ª safra), Distrito Federal, Goiás, Paraíba e Paraná (3ª safra) – sendo Alagoas, Bahia, Ceará, Mato Grosso, Pará, São Paulo, Sergipe e Tocantins (colheita até outubro 3ª safra); Pernambuco (colheita até novembro 3ª safra); Mato Grosso do Sul e Minas Gerais (colheita até dezembro 3ª safra); Tocantins e Espírito Santo (plantio a partir de novembro 1ª safra); Maranhão e Piauí (plantio a partir de dezembro 1ª safra); Paraná (plantio a partir de dezembro 2ª safra); Ceará, Distrito Federal, Goiás, Minas Gerais, Paraíba, Piauí, Santa Catarina, São Paulo e Rio Grande do Sul (plantio a partir janeiro 2ª safra); e Acre, Espírito Santo, Mato Grosso, Mato Grosso do Sul, Pernambuco, Rio de Janeiro, Rio Grande do Norte, Rondônia e Tocantins (plantio a partir de fevereiro 2ª safra);

a.3) Milho em todas as Regiões Estados analisados, sendo Roraima (colheita até outubro 1ª safra e novembro 2ª safra); Alagoas, Bahia e Sergipe (colheita até dezembro 2ª safra); Tocantins e Maranhão (plantio a partir de novembro e fevereiro); Piauí (plantio a partir de dezembro e fevereiro); Ceará, Distrito Federal, Goiás, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Paraíba, Paraná e Pernambuco (plantio a partir de janeiro 2ª safra); Rio Grande do Norte (plantio a partir de fevereiro 1ª safra); e Maranhão, Piauí, Rondônia, Roraima, São Paulo e Tocantins (plantio a partir de fevereiro 2ª safra);

a.4) Soja em todas as Regiões e Estados analisados, sendo Roraima (colheita até outubro) e Pará (plantio a partir de janeiro); e

a.5) Trigo somente nas Regiões Sudeste e Sul, sendo São Paulo (colheita até outubro); Rio Grande do Sul (colheita até novembro); e Paraná e Santa Catarina (colheita até dezembro).

Durante a coleta do CA 2006 (maio a setembro de 2007) encontravam-se no período entre o plantio e a colheita as seguintes lavouras:

a.6) Arroz no Amazonas, Goiás, Mato Grosso, Paraná, Rio Grande do Sul, Roraima, Santa Catarina, São Paulo e Tocantins (colheita até maio); Bahia, Espírito Santo, Maranhão, Minas Gerais, Piauí e Rio de Janeiro (colheita até junho); Paraíba (colheita até julho); Ceará (colheita até agosto); Pernambuco e Rio Grande do Norte (colheita até setembro); Mato Grosso do Sul, Paraná e Santa Catarina (plantio a partir de agosto) e Rio Grande do Sul e São Paulo (plantio a partir de setembro);

a.7) Feijões exceto região Centro-Oeste (1ª safra), sendo Tocantins e Maranhão (colheita até maio 1ª safra); Distrito Federal e Rio Grande do Sul (colheita até maio 2ª safra); Santa Catarina (colheita até junho 1ª safra); Paraná, Rio Grande do Sul e São Paulo (plantio a partir de agosto 1ª safra); e Santa Catarina (plantio a partir de setembro 1ª safra);

a.8) Milho em todas as Regiões e safras – exceto Distrito Federal (1ª safra) – sendo Acre, Amapá, Espírito Santo, Mato Grosso do Sul, Paraná, Rio de Janeiro, Rondônia e São Paulo (colheita até maio 1ª safra); Goiás, Mato Grosso, Minas Gerais, Pará, Santa Catarina e Tocantins (colheita até junho 1ª safra); Bahia, Ceará, Maranhão, Pernambuco e Piauí (colheita até agosto 1ª safra); Minas Gerais e Santa Catarina (plantio a partir de agosto 1ª safra) e Mato Grosso do Sul e Paraná (plantio a partir de setembro 1ª safra);

a.9) Soja em Roraima; Minas Gerais, Rio Grande do Sul, Santa Catarina, São Paulo e Tocantins (colheita até maio); Piauí (colheita até junho); Maranhão

(colheita até julho); Pará (colheita até agosto); Mato Grosso, Mato Grosso do Sul, Paraná e São Paulo (plantio a partir de setembro); e

a.10) Trigo em todas as Regiões e Estados.

Ou seja, os períodos de coleta de dados dos dois censos mais recentes coincidem com os períodos de safra das lavouras selecionadas em todas as regiões do País, exceto as lavouras de trigo nas Regiões Sudeste, Centro-Oeste e Nordeste no CA 2017 e as lavouras de feijão (1ª safra) no Centro-Oeste no CA 2006.

Considerando que essas lavouras não abrangidas nos períodos de coleta têm pequena participação nos totais nacionais e regionais, as informações obtidas sugerem uma possibilidade de ampla cobertura dos estabelecimentos agropecuários, inclusive aqueles com posse precária.

Em nível nacional, a boa abrangência da cobertura do CA 2017 em termos de unidades de produção ativas também é evidenciada por outros levantamentos do IBGE quanto à declaração de residência do(a) produtor(a) no estabelecimento agropecuário e ao monitoramento das alterações do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE), que contém os endereços e as coordenadas geográficas das unidades de produção.

Outro indicador da cobertura do Censo é o número de estabelecimentos agropecuários “com área” recenseados. Nesse caso, é possível identificar particularidades regionais quanto a esse quesito.

Tabela 4 - Número de estabelecimentos agropecuários “com área” recenseados. Brasil e Unidades da Federação. Brasil e Unidade da Federação. Em unidades

UF	2017	2006	Diferença
Brasil	4.996.287	4.920.617	75.670
Rio Grande do Sul	364.193	434.615	-70.422
Paraná	304.221	362.231	-58.010
São Paulo	187.848	225.450	-37.602
Maranhão	202.276	228.055	-25.779
Alagoas	97.786	117.792	-20.006
Rio Grande do Norte	62.893	78.674	-15.781
Santa Catarina	182.489	189.546	-7.057
Sergipe	92.742	98.361	-5.619
Pernambuco	279.437	285.045	-5.608
Distrito Federal	5.240	3.953	1.287
Paraíba	161.656	160.052	1.604

Rondônia	91.132	86.164	4.968
Amapá	8.316	3.088	5.228
Mato Grosso do Sul	70.962	64.564	6.398
Mato Grosso	118.433	111.971	6.462
Roraima	16.628	9.865	6.763
Tocantins	63.039	55.626	7.413
Rio de Janeiro	65.091	56.581	8.510
Acre	37.085	27.608	9.477
Bahia	756.822	742.195	14.627
Piauí	237.272	221.300	15.972
Goiás	151.906	133.579	18.327
Amazonas	77.840	56.335	21.505
Espírito Santo	107.386	83.763	23.623
Ceará	370.717	341.482	29.235
Minas Gerais	605.304	536.786	68.518
Pará	277.573	205.936	71.637

Fonte - IBGE- Censo Agropecuário (tabela especial 3 e tabela 6878 do Sidra). Elaboração própria.

A partir da Tabela 4 é possível observar que o número de estabelecimentos “com área” recenseado no País manteve-se estável (+ 75,7 mil unidades, ou cerca de 1,5%). No entanto, essa proporção foi consideravelmente maior na Região Norte (+ 127 mil unidades, ou cerca de + 28,6%) e menor na Região Sul (- 135,5 mil unidades, ou – 13,7%). Pelo exposto anteriormente, essas variações estão pouco associadas ao calendário entre o plantio e a colheita das lavouras selecionadas. Os estabelecimentos “sem área” são analisados na seção 2d deste texto.

b) Data de referência. Os possíveis impactos da mudança da data de referência podem ser analisados sob dois principais aspectos: o número de estabelecimentos e o número de pessoas ocupadas nessas unidades de produção:

b.1) Número de estabelecimentos. No caso do número de estabelecimentos, optou-se pela comparação entre as unidades produtivas (total e familiares) com lavouras que na data de referência do Censo Agropecuário 2017 (31 de setembro) estavam em período de entressafra, ou seja, com menor probabilidade de serem representados nos dados censitários conforme a hipótese apresentada na introdução desse texto.

As tabelas a seguir mostram as lavouras selecionadas que se encontravam fora do período entre o plantio e a colheita, em cada Unidade da Federação:

b.1.1) Arroz no Acre, Alagoas, Amapá, Bahia, Ceará, Espírito Santo, Goiás, Maranhão, Minas Gerais, Mato Grosso, Pará, Paraíba, Piauí, Rio de Janeiro, Rondônia e Tocantins;

b.1.2) Feijão na Bahia (3ª safra), Distrito Federal (3ª safra), Espírito Santo, Goiás (3ª safra), Maranhão, Mato Grosso (3ª safra), Minas Gerais (3ª safra), Piauí, Rio de Janeiro e Tocantins (3ª safra);

b.1.3) Milho no Acre, Amapá, Bahia (2ª safra), Ceará, Distrito Federal, Espírito Santo, Goiás, Maranhão, Mato Grosso, Pará, Pernambuco (2ª safra), Piauí, Rio de Janeiro, São Paulo, Sergipe e Tocantins;

b.1.4) Soja na Bahia, Distrito Federal, Goiás, Maranhão, Minas Gerais, Pará, Piauí, Rio Grande do Sul, Rondônia, Santa Catarina e Tocantins; e

b.1.5) Trigo (em nenhum estado).

Tabela 5 - Diferença entre o número total de estabelecimentos agropecuários e o número total de estabelecimentos familiares (AF) com cultivo de arroz, feijões, milho e soja entre 2017 (entressafra) e 2006, por Unidades da Federação. Em unidades.

	Arroz		Feijões		Milho		Soja	
UF	Total	AF	Total	AF	Total	AF	Total	AF
BA	-4.117	-3.794	-8.775	-16.404	-101.233	-101.724	25	-42
MG	-17.427	-15.236	-13.283	-15.657	-39.758	-40.046	3.263	1.140
MA	-63.327	-63.845	27.417	22.247	-5.628	-12.465	172	-1
CE	-22.630	-21.457	0	0	15.844	-24.618	..	..
PE	..	..	..	..	-38.924	-40.692	..	..
PI	-34.983	-34.312	51.656	32.040	-8.582	-19.658	89	-5
SE	..	..	..	..	-16.629	-16.382	..	..
RS	..	..	..	..	..	..	-9.604	-14.741
RO	-11.685	-10.917	0	0	..	..	151	-61
PA	-12.056	-11.303	0	0	3.112	735	491	112
MT	-4.777	-4.320	-895	-912	-779	-2.737	..	..
GO	-7.669	-6.169	-234	-422	-2.036	-2.651	3.255	1.401
TO	-8.903	-7.487	4.408	3.457	-207	-708	710	134
SP	..	..	..	..	-5.808	-4.178	..	..
PB	-4.181	-3.798	0	0	..	..	..	..
RJ	-529	-396	46	-86	-629	-639	..	..

AL	-327	-340	0	0	..	..	..	..
AC	-3.688	-3.393	0	0	4.146	3.392	..	..
DF	..	..	-144	-68	367	205	75	15
ES	-1.356	-1.149	3.837	2.501	-486	-901	..	..
AP	104	90	0	0	772	652	..	..
SC	..	..	..	..	..	..	6.989	5.351

.. - Não se aplica ao critério selecionado (entressafra em 2017)
0 - Dado numérico igual a zero resultante de arredondamento.
Fonte - IBGE- Censo AgropecuárioElaboração própria a partir das tabelas especiais 10, 12, 13, 14, 17 e 18 e da tabela 6957 do Sidra

As tabelas a seguir comparam os resultados dos censos mais recentes das lavouras que na data de referência do CA 2006 (31 de dezembro) estavam em período de entressafra:

b.1.6) Arroz no Ceará, Paraíba, Pernambuco, Rio Grande do Norte e Roraima;

b.1.7) Feijão no Acre, Amapá, Ceará, Paraíba, Pernambuco, Rio Grande do Norte, Rondônia e Sergipe;

b.1.8) Milho no Ceará, Paraíba, Pernambuco, Rio Grande do Norte e Roraima;

b.1.9) Soja no Pará e em Roraima; e

b.1.10) Trigo na Bahia, Distrito Federal, Goiás, Mato Grosso do Sul, Minas Gerais, Rio Grande do Sul e São Paulo.

Tabela 6 - Diferença entre o número total de estabelecimentos agropecuários e o número total de estabelecimentos familiares (AF) com cultivo de arroz, feijões, milho, soja e trigo entre 2017 e 2006 (entressafra). Unidades da Federação. Em unidades.

	Arroz		Feijões		Milho		Soja		Trigo	
	Total	AF	Total	AF	Total	AF	Total	AF	Total	AF
PE	-811	-761	-10.250	-15.625	-38.924	-40.692	..	..	..	..
CE	-22.630	-21.457	45.146	-525	15.844	-24.618	..	..	..	..
PB	-4.181	-3.798	-2.980	-9.080	-20.550	-24.803	..	..	..	..
RN	-1.487	-1.242	-14.585	-13.578	-8.426	-7.893	..	..	..	..
RO	..	..	-10.209	-9.739	..	..	..	..	..	..
SE	..	..	-6.579	-6.677	..	..	..	..	..	..
RS	..	..	..	..	..	..	..	..	-50	-773
AC	..	..	-815	-704	..	..	..	..	..	..
RR	-898	-874	-815	-704	1.514	1.249	49	12	..	..
MS	..	..	..	..	..	..	..	..	-83	-16

BA	..	..	..	..	..	..	..	..	3	0
DF	..	..	..	..	..	..	..	..	-3	0
GO	..	..	..	..	..	..	..	..	9	3
AP	..	..	19	16	..	..	..	..	..	..
MG	..	..	..	..	..	..	..	..	318	36
PA	..	..	..	..	..	..	491	112	..	..
SP	..	..	..	..	..	..	..	..	591	240

.. - Não se aplica ao critério selecionado (entressafra em 2006)
0 - Dado numérico igual a zero resultante de arredondamento.
Fonte - IBGE- Censo Agropecuário (tabelas especiais 10, 12, 13, 14, 17 e 18 e da tabela 6957 do Sidra). Elaboração própria.

A comparação entre os dados das Tabelas 5 e 6 permite verificar que o número de lavouras em situação de entressafra na data de referência do CA 2017 é expressivamente maior, mais abrangente e mais disperso do que no período análogo do CA 2006. Isto sugere que a mudança da data de referência do censo pode estar associada à redução do número total de estabelecimentos e do número total de estabelecimentos familiares. As principais exceções são os estabelecimentos familiares com cultivos de feijão fradinho no Piauí e no Maranhão, que aumentaram consideravelmente.

b.2) Pessoal Ocupado. A tabela a seguir mostra o número de pessoas ocupadas na data de referência dos dois censos mais recentes.

Tabela 7 - Diferença do número de pessoas ocupadas nos estabelecimentos agropecuários (total e familiares) nas datas de referência do Censo Agropecuário 2017 e do Censo Agropecuário 2006. Brasil e Unidades da Federação. Em unidades.

UF	Pessoal Ocupado Total	Pessoal Ocupado AF
Brasil	-1.463.080	-2.165.986
Bahia	-220.310	-324.883
Maranhão	-298.730	-320.487
Ceará	-217.344	-281.272
Rio Grande do Sul	-239.412	-273.020
Paraná	-270.456	-243.936
Piauí	-161.506	-199.259
Pernambuco	-165.182	-189.326
Minas Gerais	-60.584	-105.881
Santa Catarina	-69.711	-101.894
Alagoas	-124.830	-96.008
Paraíba	-66.201	-94.801
Sergipe	-34.639	-54.756
São Paulo	-77.653	-50.560
Rio Grande do Norte	-33.632	-46.155
Rondônia	-6.945	-27.658

Rio de Janeiro	2.875	216
Tocantins	27.599	714
Distrito Federal	-533	2.658
Mato Grosso do Sul	43.778	5.949
Mato Grosso	64.117	6.747
Espírito Santo	39.690	7.103
Amapá	18.003	14.149
Goiás	72.541	16.920
Acre	26.935	17.184
Roraima	37.561	26.484
Amazonas	64.052	34.068
Pará	187.437	111.718

Fonte - IBGE- Censo Agropecuário (tabela especial 5 e tabela 6884 do Sidra). Elaboração própria

A Tabela 7 mostra que entre os dois censos mais recentes houve uma importante redução do número de pessoas ocupadas nos estabelecimentos familiares na Região Nordeste e na Região Sul. No entanto, somente na Região Nordeste e no Rio Grande do Sul isso pode ser eventualmente associado às lavouras selecionadas em período de entressafra. Em sentido contrário, os estados da Região Norte (exceto Rondônia) apresentaram crescimento no número de pessoas ocupadas em estabelecimentos familiares, mesmo com algumas das lavouras selecionadas estando em período de entressafra nas datas de referência dos censos. O tema da ocupação nos estabelecimentos é retomado na seção 3.3 desse texto.

c) Período de referência. Os possíveis impactos da mudança no período de referência podem ser analisados sob dois aspectos principais: as alterações das condições climáticas e dos Valores Brutos da Produção (VBP).

c.1) Condições climáticas

No período de referência do CA 2017 (1º de outubro de 2016 a 31 de setembro de 2017) a Conab avaliou as condições climáticas como favoráveis às atividades agrícolas. As precipitações estiveram dentro da faixa normal nas principais regiões produtoras de grãos, tanto para as lavouras de primeira safra quanto nas demais.

Algumas lavouras e localidades foram pontualmente atingidas por irregularidades das chuvas, principalmente no norte de Minas Gerais (feijão fradinho 1ª safra), na Paraíba (feijão de cor e feijão fradinho 2ª safra; milho 1ª safra); em Pernambuco (feijão fradinho 2ª safra; milho 1ª safra) e no centro sul e no centro norte da Bahia (feijão de cor 1ª safra). O excesso de chuvas afetou o Paraná (feijão de cor 2ª safra); o Rio Grande do Sul (feijão preto 1ª e 2ª safra); o

Mato Grosso (feijão fradinho 2ª safra) e Alagoas (feijão de cor e feijão fradinho 3ª safra; milho 2ª safra). As geadas impactaram localmente a produção no Paraná (trigo) e no Mato Grosso do Sul (trigo).

As informações e a avaliação da Conab das condições climáticas sobre as lavouras selecionadas no período de referência do Censo Agropecuário 2006 (1º de janeiro a 31 de dezembro de 2006) não estão disponíveis em meio digital. Os dados do mostram que o estado do Rio Grande do Sul e localidades do Ceará, Bahia, Paraná, São Paulo e Santa Catarina tiveram prolongados períodos com anomalias de precipitação, atingindo as lavouras de milho, trigo e feijões, nesta ordem de importância.

c.2) Valor Bruto da Produção

A Tabela 8 a seguir compara o VBP obtido em estabelecimentos familiares nas lavouras selecionadas nos dois censos mais recentes.

Tabela 8 - Diferença do Valor Bruto de Produção de lavouras selecionadas em estabelecimentos familiares entre 2017 e 2006. Brasil e Unidades da Federação. Em 1.000.000 de reais nominais.

UF	Arroz	Feijão preto	Feijão de cor	Feijão fradinho	Milho	Soja	Trigo
Brasil	-454	-35	-239	-459	451	6.669	272
Ceará	-44	-13	-40	-132	-299	0	0
Maranhão	-365	-2	-17	-4	-103	1	0
Bahia	-4	-8	-70	-126	-115	-2	0
Pernambuco	-5	-13	-18	-50	-100	0	0
Piauí	-74	0	0	-19	-12	0	0
Paraíba	-14	-2	-11	-41	-34	0	0
Alagoas	-1	0	-67	-9	-21	0	0
Pará	-76	-1	-9	-13	-22	33	0
Rio Grande do Norte	-4	0	-4	-30	-24	0	0
Minas Gerais	-24	2	23	-9	-231	200	2
Rondônia	-19	0	-4	-12	-5	14	0
Acre	-20	0	-4	-5	9	0	0
Rio de Janeiro	-2	-2	0	0	-4	0	0
Amazonas	-2	0	0	-5	1	0	0
Roraima	-4	0	0	0	0	2	0
Sergipe	10	0	-4	-5	-3	0	0
Amapá	0	0	0	0	1	0	0
Distrito Federal	0	0	0	0	1	0	0
Espírito Santo	-1	5	4	0	12	0	0

Tocantins	-21	-1	-1	3	-2	46	0
Mato Grosso do Sul	-6	-1	-3	0	68	195	0
Goiás	-21	0	5	1	55	250	0
São Paulo	-2	0	6	-7	64	227	8
Mato Grosso	0	0	0	7	114	441	0
Santa Catarina	138	-16	-8	-1	51	457	8
Paraná	-3	11	-4	-3	585	1.932	55
Rio Grande do Sul	110	8	-11	0	464	2.872	200

Fonte - IBGE- Censo Agropecuário (tabelas especiais 10, 12, 13, 14, 17, 18 e 19 e tabela 6957 do Sidra). Elaboração própria.

A Tabela 8 mostra que houve uma enorme diminuição nominal do Valor Bruto da Produção das lavouras selecionadas nos estabelecimentos familiares, especialmente na Região Nordeste. No sentido oposto, houve um significativo aumento desse Valor nas unidades familiares da Região Sul e da Região Centro-Oeste, especialmente as que cultivaram soja no Rio Grande do Sul e no Paraná.

Os Cadernos Estatísticos da Conab disponíveis em meio digital não permitem aferir a contribuição dos preços unitários na formação do VBP das lavouras selecionadas em 2006 e 2017. As Tabelas 11 e 12 a seguir permitem dimensionar a contribuição da área plantada e da área colhida no cálculo desse Valor.

d) Produção/criação de empregados/moradores em área do estabelecimento agropecuário. O fato do CA 2017 não abrir o questionário de novo estabelecimento para o(a) produtor(a) empregado(a)/morador(a) em área sujeita à administração do(a) produtor(a)/proprietário(a), trouxe como resultado esperado a redução no número de estabelecimentos de produtores “sem área” em relação ao censo anterior.

A tabela a seguir sugere que a alteração da forma de captação dos produtores “sem área” pode estar mais associada à redução do número de estabelecimentos familiares do que do conjunto de estabelecimentos. Os estados da Região Nordeste concentraram a maior parte desta diminuição.

Tabela 9 - Diferença do número de total de estabelecimentos e dos estabelecimentos familiares (AF) sem área entre 2006 e 2017. Brasil e Unidades da Federação. Em unidades.

UF	Total	Sem área	AF	AF Sem área
Brasil	-102.312	-177.982	-407.697	-181.676
Maranhão	-67.274	-41.495	-75.958	-43.259
Ceará	13.313	-15.922	-40.241	-22.864
Piauí	223	-15.749	-20.963	-16.817

Pernambuco	-23.102	-17.494	-36.159	-16.015
Bahia	1.290	-13.337	-54.552	-13.253
Pará	59.670	-11.967	44.142	-11.750
Minas Gerais	55.936	-12.582	9.217	-11.257
Amazonas	14.175	-7.330	8.980	-7.432
Paraná	-65.909	-7.899	-69.838	-6.727
Paraíba	-4.068	-5.672	-19.823	-5.266
Rio Grande do Sul	-76.378	-5.956	-81.108	-5.155
Alagoas	-24.790	-4.784	-27.224	-4.428
Rio Grande do Norte	-19.601	-3.820	-19.732	-3.611
Santa Catarina	-10.602	-3.545	-22.594	-3.266
Sergipe	-7.332	-1.713	-16.746	-1.732
Rio de Janeiro	6.731	-1.779	286	-1.730
Goiás	16.482	-1.845	8.392	-1.680
Acre	7.873	-1.604	5.890	-1.491
São Paulo	-39.002	-1.400	-26.030	-1.383
Mato Grosso	5.692	-770	-1.396	-794
Rondônia	4.360	-608	-128	-553
Tocantins	7.241	-172	2.297	-389
Espírito Santo	23.653	30	13.241	-256
Roraima	6.536	-227	4.399	-249
Amapá	4.980	-248	4.082	-223
Mato Grosso do Sul	6.300	-98	2.969	-101
Distrito Federal	1.291	4	900	5

Fonte - IBGE- Censo Agropecuário (tabelas especiais 3 e da tabela 6878 do Sidra). Elaboração própria.

A tabela a seguir caracteriza os estabelecimentos “sem área” pelo tipo de produtor(a) no CA 2017.

Tabela 10 - Estabelecimentos agropecuários sem área segundo o tipo de produtor “sem área” no Brasil e na Região Nordeste, Pará e Amazonas. 2017. Em unidades.

Tipo de produtor sem área	Brasil	NE (A)	PA (B)	AM (C)	(A)+(B)+(C)/ Brasil
Total estabelecimentos produtor sem área	77.037	61.118	4.126	3.119	88,7%
Terras arrendadas, em parceria ou ocupadas	34.125	32.305	467	235	96,7%
Outra situação	17.154	10.890	1.595	1.035	78,8%
Extrativista	10.288	6.776	1.796	667	89,8%
Vazantes rios, roças itinerantes, beira de estradas	7.066	5.627	143	1.148	97,9%
Criador de animais em beira de estrada	4.936	4.505	65	19	93,0%
Produtor(a) de mel	3.468	1.015	60	15	31,4%

Fonte - IBGE- Censo Agropecuário (tabulação especial não publicada).

3. OUTROS CONDICIONANTES

A leitura dos levantamentos de acompanhamento de safra elaborados pela Conab sugere que além da análise pontual das condições vigentes no período de referência do CA 2017, é importante destacar os processos decorrentes de variações climáticas nos anos anteriores a ele e suas implicações sobre o financiamento das lavouras, a intenção de plantio e as práticas utilizadas. Essa análise é fundamental para a compreensão dos casos das Regiões Norte e Nordeste.

A partir do último trimestre de 2015 o padrão de chuvas em grande parte do Brasil foi influenciado pelos efeitos do fenômeno El Niño de características muito fortes. Na região Amazônica, as precipitações da verificadas na estação chuvosa diminuíram cerca de 50% em relação à média e continuaram abaixo da média até agosto de 2016. Essa intensidade de redução das chuvas não era registrado desde 2002. A redução da precipitação em praticamente toda a região, com déficit acentuado de chuva envolvendo toda a região e até sem registro de chuvas durante 90 dias, como aconteceu no Acre, esteve associada ao aumento de incêndios florestais; ao registro de altas temperaturas máximas variando acima da média; e a situações de emergência pública, inclusive em diversas cidades próximas a capitais dos estados, que decretaram situação de alerta ou de calamidade pública em razão de problemas de abastecimento de água.

Os prejuízos para as atividades agropecuárias ocorreram especialmente nas lavouras irrigadas (fruticultura e hortaliças) e na produção leiteira. Em razão

de vazantes atípicas nos principais rios e bacias da Amazônia que atuam como corredores logísticos do país, houve restrição prolongada do escoamento de parte da produção agrícola, principalmente soja e milho de Mato Grosso e Rondônia, bem como de combustíveis e fertilizantes, com destino a Porto Velho e Manaus.

Os cenários de seca extrema a excepcional foram observados em todos os estados do Nordeste, sendo que Paraíba, Ceará e Rio Grande do Norte, apresentaram os maiores números de municípios que decretaram situação de emergência. A intensidade da seca em 2016 apresentou uma característica de seca grave no início do ano e em março, evoluiu com máximo em outubro, o mesmo observado para a intensidade de seca extrema, enquanto a seca excepcional se tornou evidente em junho (Inmet, 2016).

No Nordeste, a safra 2015/2016 foi antecedida por um o período prolongado de seca e, consequentemente, por um acúmulo de prejuízos. O evento El Niño nessa safra agravou a situação de seca plurianual iniciada em 2012, tornando-a o período mais crítico em termos de totais de chuva desde 1911. Não por acaso, é chamada de a “Grande Seca” (Aquino & Nascimento, 2019).

Os sinais de seca começaram a aparecer em dezembro de 2011 e se intensificaram durante o verão e outono de 2012, gerando deficiência hídrica em quase todo o semiárido até 2014, desde o centro-sul da Bahia até o Rio Grande do Norte e o Ceará. Entre 2013 e 2015 a maior concentração de déficit hídrico incluiu particularmente o norte da Bahia, oeste do Pernambuco e o leste do Piauí (Marengo et al., 2016)

Tabela 11 - Área plantada e diferença entre área plantada e área colhida em lavouras selecionadas. Região Nordeste. Vários anos. Em 1.000 ha.

	Arroz		Feijão		Milho		Soja	
Ano	Área plantada	Dif.	Área plantada	Dif.	Área plantada	Dif.	Área plantada	Dif.
2006	735	-19	2.348	-173	2.868	-145	1.488	0
2012	603	-21	1.471	-452	2.462	-684	2.115	0
2013	577	-28	1.361	-184	2.272	-300	2.327	-20
2014	531	-2	1.737	-196	2.820	-323	2.581	0
2015	357	-15	1.629	-246	2.688	-358	2.870	-2
2016	267	-24	1.456	-309	2.459	-520	2.884	-1
2017	242	-1	1.543	-238	2.649	-275	3.097	0

Fonte - IBGE – Produção Agrícola Municipal (tabela 1612 do Sidra). Elaboração própria.

A partir da Tabela 11 é possível observar na Região Nordeste uma acentuada redução da área plantada com arroz, além de relevantes e persistentes perdas de área até a colheita das lavouras selecionadas em todos os anos do período de 2012 a 2017, exceto a soja, que praticamente teve a sua área plantada dobrada e perdas mínimas de área até a colheita. As diferenças de área plantada e de área colhida são fatores importantes para a compreensão da redução do Valor Bruto de Produção, já mencionada.

A tabela a seguir permite distinguir a particularidade do comportamento da área plantada e da área colhida das lavouras selecionadas na Região Nordeste que pode ser fortemente associado ao período da Grande Seca.

Tabela 12 - Área plantada e diferença entre área plantada e área colhida em lavouras selecionadas. Região Sul. Vários anos. Em 1.000 ha

	Arroz		Feijão		Milho		Soja		Trigo	
Ano	Área plantada	Dif.	Área plantada	Dif.	Área plantada	Dif.	Área plantada	Dif.	Área plantada	Dif.
2006	1.238	0	851	-11	4.685	-127	8.132	-5	1.647	-210
2012	1.227	-5	645	-12	4.656	-134	9.178	-114	1.850	-29
2013	1.268	-2	634	-15	4.532	-55	10.012	-1	2.136	-136
2014	1.294	-1	678	-4	3.928	-9	10.561	-4	2.665	-1
2015	1.304	-6	566	-2	3.699	-2	11.113	-2	2.278	-17
2016	1.262	-26	521	-10	3.685	-9	11.580	-31	1.945	-1
2017	1.277	-3	584	-22	4.032	-2	11.447	-11	1.687	-7

Fonte - IBGE- Pesquisa Agrícola Municipal (tabela 1612 do Sidra), Elaboração própria.

A Tabela 12 revela que, com exceção do ano agrícola 2016/2017, houve um aumento continuado na área plantada com soja, uma redução da área com feijões e uma relativa estabilidade nas áreas plantadas com arroz na Região Sul no período entre os censos mais recentes. A área plantada com milho diminui acentuadamente após 2013 e a área com trigo aumenta entre 2013 e 2015 mas retorna ao mesmo patamar de 2006 no ano agrícola de apuração do censo mais recente. As diferenças nominais entre as áreas plantadas e as áreas colhidas são maiores no milho, no trigo e na soja, mas não constituem uma sequência de repetições. As diferenças relativas são maiores nas áreas plantadas com feijões.

É válido supor efeitos da seca prolongada sobre a pecuária, mesmo que seja difícil quantificá-los. A tabela a seguir procura caracterizar a evolução do número de estabelecimentos familiares com as principais criações. Ela complementa a Tabela 3, que indica uma possível migração da atividade da produção de lavouras temporárias para a da pecuária na maioria das Unidades da Federação da Região Nordeste. Este complemento é relevante para a análise

da variação do número de pessoas ocupadas, já abordado em outra seção deste texto, e da maior pressão por obtenção de rendas fora do estabelecimento.

Tabela 13 - Diferença do número de estabelecimentos agropecuários familiares entre 2006 e 2017, com criações selecionadas. Brasil e Grandes Regiões. Em 1.000 unidades.

Região	Bovinos	Caprinos	Ovinos	Suínos	Galinhas, galos, frangas, frangos e pintos
Brasil	-230,8	13,2	43,1	-109,5	-30,8
Norte	20,8	2,3	-0,1	29,3	63,0
Nordeste	-129,9	14,9	40,9	-52,4	-20,8
Sudeste	-6,5	-0,5	-2,8	-20,5	8,5
Sul	-135,2	-4,3	5,4	-81,2	-99,0
Centro-Oeste	19,9	0,8	-0,3	15,2	17,5

Fonte - IBGE – Censo Agropecuário (tabelas especiais 25, 27, 29, 31 e 33 e tabela 6907 do Sidra). Elaboração própria.

A partir da Tabela 13 é possível observar que no período entre os dois censos mais recentes houve uma intensa mudança nas criações existentes nos estabelecimentos familiares, especialmente do rebanho bovino nas Regiões Sul e Nordeste. No caso da Região Nordeste, a redução do número de estabelecimentos com bovinos ocorre simultaneamente ao aumento do número de estabelecimentos com ovinos e caprinos, ainda que em menor proporção. Isso sugere uma mudança interna ao grupo de atividade da pecuária e outras criações. No caso da Região Sul, essa diminuição também ocorre de forma expressiva nos rebanhos de galinhas, frangos e assemelhados, de suínos e de caprinos.

Outro aspecto importante para compreensão dos resultados do CA 2017 está relacionado ao período de referência é a edição do Acórdão 775/2016, de abril de 2016, e do Acórdão nº 2.451/2016 de setembro de 2016, ambos do Plenário do Tribunal de Contas da União (TCU), que determinaram por medida cautelar a suspensão dos processos de seleção de novos beneficiários para a reforma agrária; de assentamento de novos beneficiários já selecionados; de novos pagamentos e de remissão de créditos da reforma agrária para os beneficiários com indícios de irregularidade; e o acesso a outros benefícios e políticas públicas concedidos em função de o(a) beneficiário(a) fazer parte do Programa Nacional de Reforma Agrária (PNRA), como os Programas Nacional de Fortalecimento da Agricultura (Pronaf), Garantia Safra, Minha Casa Minha Vida – Habitação Rural, Bolsa Verde e de Assistência Técnica e Extensão Rural, entre outros. A medida cautelar foi revogada em setembro de 2017, por meio do

Acórdão/TCU/nº 1976/2017-Plenário. Ou seja, durante todo o período de referência do CA 2017 (01/10/2016 a 30/09/2017) é possível estimar que cerca de 88 mil beneficiários(as) do PNRA não teve acesso a novos recursos decorrentes das políticas públicas federais específicas e outra parte não dimensionável não teve acesso à terra ou não teve acesso à terra na condição de concessionário e assentado(a) aguardando titulação

4. CONSIDERAÇÕES FINAIS

A análise dos condicionantes do CA 2017 aqui abordados sugere possibilidades de associação da diminuição do número de estabelecimentos agropecuários familiares em relação ao censo anterior pela combinação de três ou mais elementos apresentados neste texto. Estas possibilidades constituem novas hipóteses a serem melhor investigadas a partir de tabulações especiais não disponibilizadas no Sidra e de outras pesquisas nacionais.

No caso da Região Nordeste, os primeiros elementos a serem destacados são a Grande Seca e a mudança metodológica para o cômputo dos estabelecimentos “sem área”. Existem indícios que ambos podem ter tido os seus efeitos acentuados pela mudança da data de referência para um período de entressafra em importantes lavouras da região.

A enorme queda no Valor Bruto da Produção agropecuária e no número de pessoas ocupadas nos estabelecimentos recenseados em 2006 e 2017 sugere a possibilidade de mudança de grupo de atividade econômica e/ou busca mais acentuada de rendas fora da unidade produtiva. Isso é compatível com a análise realizada por Joacir Aquino e Carlos Nascimento, com base nos dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) entre 2011 e 2015, que revelam a distribuição da população rural ocupada em atividades agropecuárias ou em atividades não-agropecuárias e a composição da renda média familiar rural, segundo o tipo de família e as diferentes fontes de renda do trabalho (agrícola e não agrícola) e do não-trabalho (aposentadorias/pensões e outras fontes) (Aquino & Nascimento, 2019).

Esta hipótese também é compatível com os dados apresentados por Mauro del Grossi e outros autores que mostram que entre pequenos produtores (com área total até quatro módulos fiscais), 362.890 estabelecimentos na Região Nordeste deixaram de ser classificados como agricultores familiares segundo o Decreto nº 9.064/2017, devido à maior importância das rendas obtidas fora das suas unidades produtivas (Del Grossi et al., 2019).

As Regiões Nordeste e Sul foram as que apresentaram a maior queda do número de pessoas ocupadas, especialmente entre as pessoas com laço de parentesco com o(a) produtor(a) em estabelecimentos familiares. Não existem indícios que isso pode ser eventualmente associado às lavouras em período de entressafra no Paraná e em Santa Catarina. Uma hipótese explicativa a ser melhor investigada recai sobre as mudanças na atividade da pecuária.

Outra hipótese a ser investigada a partir da utilização de dados do CA 2006 segundo as normas de classificação da Agricultura Familiar vigentes em 2017 é a relação entre a diminuição do número de pessoas ocupadas com laço de parentesco com o(a) produtor(a), o aumento da mão de obra contratada e a idade do(a) produtor(a). Os dados apresentados por Mauro del Grossi e outros autores revelam que os pequenos estabelecimentos (66.196 unidades na Região Nordeste e 18.028 unidades na Região Sul) deixaram de ser classificados como agricultores(as) familiares devido ao predomínio da mão de obra contratada (Del Grossi et al., 2019).

Na Região Sul existem menos evidências de possíveis associações entre os resultados obtidos e as mudanças nas metodologias adotadas nos dois censos mais recentes, inclusive quanto ao número de estabelecimentos “sem área”. Assim, outros condicionantes que escapam ao escopo deste texto parecem ter maior poder explicativo sobre as variações observadas.

Um deles diz respeito à origem das rendas obtidas. Os cálculos disponíveis para a Região Sul revelam que 85.728 estabelecimentos deixaram de ser classificados como de agricultores(as) familiares segundo o Decreto nº 9.064/2017, devido à maior importância das rendas obtidas fora das unidades produtivas (Del Grossi et al., 2019). Os dados disponíveis mostram que entre as lavouras pesquisadas, o Valor Bruto da Produção somente apresentou variação negativa relevante nos casos do feijão preto e do feijão de cor em Santa Catarina e do feijão de cor no Rio Grande do Sul.

Na Região Sul, as tabelas disponíveis no Sidra sugerem que os estabelecimentos familiares que devam requerer investigações específicas sobre a diminuição do seu número entre os dois censos mais recentes são os que cultivaram milho (- 131,2 mil estabelecimentos), feijão preto (- 24,6 mil) e feijão de cor (- 20,4 mil) em condição precária de posse da terra (arrendatários, parceiros, ocupantes/comodatários). Esta hipóteses necessitam ser testadas a partir da utilização de dados do CA 2006 segundo as normas de classificação da Agricultura Familiar vigentes em 2017.

Entre as demais hipóteses a serem melhor investigadas a partir da utilização de dados do CA 2006 segundo as normas de classificação da Agricultura Familiar vigentes em 2017 estão a evolução de lavouras concentradas territorialmente não analisadas neste texto, como o fumo na Região Sul; a idade do(a) produtor(a); e a dinâmica particular empreendida no território do Matopiba, compreende 337 municípios do Maranhão, Tocantins, Piauí e Bahia, predominantemente no Bioma Cerrado (Favareto et al., 2019).

5. REFERÊNCIAS BIBLIOGRÁFICAS

Aquino, J. R. A. & Nascimento, C. A. (2019). A Grande Seca e as fontes de ocupação e renda das famílias rurais no Nordeste do Brasil (2011-2015). 57º Congresso da SOBER, Ilhéus/BA, 21 a 25 de julho de 2019.

Companhia Nacional de Abastecimento (Conab). Safra 2016/17. Primeiro levantamento. Acompanhamento safra brasileira de grãos, v. 4, n. 1, p. 1-162. Brasília: Conab, outubro de 2016.

Companhia Nacional de Abastecimento (Conab). (2017). Safra 2016/17. Décimo segundo levantamento. Acompanhamento safra brasileira de grãos, v. 4, n. 12, p. 1-158. Brasília: (Conab), setembro de 2017.

Companhia Nacional de Abastecimento (Conab). Calendário de plantio e colheita de grãos no Brasil 2019. Brasília: (Conab), 2015. 75 p.

Del Grossi, M.E. (2019). Algoritmo para delimitação da Agricultura Familiar no Censo Agropecuário 2017, visando a inclusão de variável no Banco de Dados do Censo, disponível para ampla consulta. Brasília: FAO; SAF/MAPA; FINATEC, jun. 2019. 25 p.

Del Grossi, M., Florido, A.C. & Rodrigues, L.F.P. (2019). Agricultura Familiar no Censo Agropecuário. Principais causas de exclusão da Agricultura Familiar nos algoritmos. Versão 8 de novembro de 2019. 3 p.

Del Grossi, M., Florido, A.C., Rodrigues, L.F.P. & Oliveira, M.S. (2019). Comunicação de Pesquisa: Delimitando a Agricultura Familiar nos Censos Agropecuários Brasileiros. Rev. NECAT, Florianópolis, v. 8, n. 16, p. 40-45.

Favareto, A. (org.), Nakagawa, L., Pó, M., Seifer, P. & Kleeb, S. (2019). Entre chapadas e baixões do Matopiba: dinâmicas territoriais e impactos socioeconômicos na fronteira da expansão agropecuária no cerrado. São Paulo: Prefixo Editorial 92545, 272p.

Hoffmann, R. (2007). Distribuição da renda e da posse da terra no Brasil. *In:* Ramos, Pedro (org.) Dimensões do agronegócio brasileiro: políticas, instituições e perspectivas. Brasília: MDA, 2007, p. 172-225 (Nead Estudos, 15).

Instituto Brasileiro de Geografia e Estatística (IBGE). Censo Agropecuário 2006. Brasil, Grandes Regiões e Unidades da Federação. Segunda Apuração. Censo Agropecuário, Rio de Janeiro, p. 1-774, 2012.

Instituto Brasileiro de Geografia e Estatística (IBGE). Censo agropecuário 2017. Resultados definitivos. Censo agropecuário, Rio de Janeiro, v. 8, p.1-105, 2019.

Instituto Brasileiro de Geografia e Estatística (IBGE). Censo Agropecuário. Estabelecimentos agropecuários levantados pelo Censo Agropecuário 2006 que atendem às regras da época (2006) e os que atendem às regras do Censo Agropecuário (2017) para enquadramento na agricultura familiar e no Pronaf. Planilhas LibreOffice Calc (agri_familiar_2006_2). 2017(b). Disponível em:

https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/21814-2017-censo-agropecuario.html?=&t=downloads.

Instituto Nacional de Meteorologia (Inmet). Situação da seca observada nas Regiões Norte e Nordeste do Brasil em 2016. Nota Técnica, Brasília, março de 2016. 8 p.

Marengo, J., Cunha, A.P. & Alves, L. M. (2016). A Seca de 2012-15 no Semiárido do Nordeste do Brasil no Contexto Histórico. Climanálise, v. 4, p. 49-54.

Matte, A. Convenções e mercados na pecuária familiar no sul do Rio Grande do Sul, Brasil. Tese (Doutorado em Desenvolvimento Rural) Universidade Federal do Rio Grande do Sul. Porto Alegre, 2017. 292 f.

AGRADECIMENTOS

Agradeço a colaboração de Antônio Carlos Simões Florido (IBGE), Caio Galvão de França, Mauro Del Grossi (UnB) e Alessandra Matte (UTFPR). A colaboração dessas pessoas não implica qualquer responsabilidade delas sobre as interpretações no presente trabalho.

Revista Brasileira de Estatística, V.80, N. 247, p.75-91, 2022

REGRESSÃO QUANTÍLICA NA ANÁLISE DA LETALIDADE DA COVID-19 NOS MUNICÍPIOS DO ESTADO DE SÃO PAULO

Gabriel T. Alves

gta1998@gmail.com

Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Helton Saulo

heltonsaulo@gmail.com

Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Rodrigo Nobre Fernandez

rodrigonobrefernandez@gmail.com

Departamento de Economia, Universidade Federal de Pelotas, Pelotas, Brasil

Leandro T. Correia

leandrotc@gmail.com

Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Jose A. Fiorucci

jafiorucci@gmail.com

Departamento de Estatística, Universidade de Brasília, Brasília, Brasil

Resumo: Neste trabalho um modelo de regressão quantílica logística para dados limitados é utilizado para analisar os determinantes da taxa de letalidade da COVID-19 nos municípios do estado de São Paulo. O modelo é utilizado ainda para fins de predição da letalidade. O uso de metodologia que leva em consideração dados limitados, como é o caso do modelo de regressão quantílica logística, é especialmente importante no contexto de regressão quantílica, uma vez que resultados mais confiáveis são obtidos com métodos que restringem a inferência dentro de um intervalo limitado, como é o caso da taxa de letalidade (limitada entre 0 e 1). Os resultados mostram que a segunda dose e dose de reforço estão associados à redução na letalidade da COVID-19. Em adição, o modelo de regressão quantílica logística teve bom desempenho preditivo.

Palavras-chave: Regressão quantílica; COVID-19; Taxa de letalidade; Predição.

Abstract: In this work, a logistic quantile regression model for bounded data is used to analyze the determinants of the case fatality rate of COVID-19 in the municipalities of the state of São Paulo, Brazil. The model is also used for lethality prediction purposes. The use of a methodology that uses bounded data, such as the quantile logistic regression model, is especially important in the context of quantile regression, since more reliable results are obtained with methods that restrict the inference within of a limited range, as is the case of the case fatality rate (limited between 0 and 1). The results show that the second and booster doses are associated with a reduction in the COVID-19 fatality rate. In addition, the logistic quantile regression model had good predictive performance.
Keywords: Quantile regression; COVID-19; Fatality rate; Prediction.

1.INTRODUÇÃO

A pandemia do novo coronavírus (COVID-19) é um tema de discussão para muitos estudos e análises. O comportamento do vírus e de pacientes que contraem a doença se tornam foco de muitas pesquisas que visam entender o retrato da situação para que no futuro decisões sejam tomadas corroboradas por dados e evidências.

Em 11 de março de 2020 a Organização Mundial da Saúde (OMS) identificou o grande número de casos gerados pelo novo Coronavírus (COVID-19) como uma pandemia que até o início de julho de 2022 já ocasionou a morte de 6 milhões e 480 mil pessoas ao redor do mundo.

Além do custo elevadíssimo em termos de vidas humanas, no que se refere ao âmbito econômico, antes e durante o surgimento das vacinas para o combate a expansão da doença, os governos ao redor do globo empregaram um conjunto de medidas que possuíam como foco combater o contágio, dentre elas: i) o funcionamento de unicamente atividades econômicas consideradas essenciais, tais quais os serviços de abastecimento de veículos, mercados/supermercados e farmácias/drogarias; ii) a recomendação do uso da máscara e do distanciamento social (EICHEINBAUM, *et. al* 2020; ALVAREZ, et al., 2020; FARBOODI, *et al.* 2020).

No intuito de verificar se as medidas restritivas são eficazes, Friedson *et al.* (2020) e Bron *et al.* (2020) utilizaram o método de controle sintético para avaliar o caso Americano e para a Suécia, respectivamente. A pesquisa elaborada por Dave et al. (2020), Di Porto et al., (2020), Fang et al., (2020) e Gupta et al. (2020) utilizam o método de diferenças-em-diferenças para avaliar se intervenções governamentais podem reduzir o número de internações ou de mortes causadas por COVID-19, contudo os resultados são contrastantes.

No contexto brasileiro, Oliveira et al. (2020) estimou que o custo econômico da intensificação das medidas de isolamento social foi de

aproximadamente R$ 844 mil por dia. Sob outro prisma, Oliveira et al. (2020) estimaram uma perda de R$ 43,36 bilhões em vendas e de arrecadação do Imposto de Circulação de Mercadorias e Serviços, algo próximo de R$ 1,56 bilhão desde os 27 primeiros dias de medidas restritivas.

Dentro deste contexto, esse estudo foca em uma medida de risco de mortalidade, a taxa de letalidade, ocasionada pelo COVID-19, em particular, que serve como instrumento de monitoramento da gravidade da pandemia. Essa taxa é calculada como a razão entre o número de mortes pelo número de casos confirmados. Para estudar-se o comportamento de taxas e proporções geralmente é utilizado o modelo de regressão beta (FERRARI e CRIBARI-NETO, 2004). Por outro lado, essa abordagem é baseada na média e pode ser inadequada se a variável dependente segue uma distribuição assimétrica, o que ocorre com a taxa de letalidade. Adicionalmente, a regressão beta, não proporciona uma visão mais abrangente da relação das variáveis explicativas sobre a variável dependente, o que pode ser observado por meio da regressão quantílica introduzida por Koenker e Basset Jr. (1978).

Em contrapartida, quando a variável dependente é limitada, podem ocorrer problemas quando métodos estatísticos tradicionais são utilizados. Em geral, usar métodos que restringem a inferência dentro do intervalo limitado leva a conclusões mais confiáveis (BOTTAI, et al. 2010).

Esse trabalho utiliza um modelo de regressão logística para dados limitados, introduzido por Bottai *et al.* (2010), para modelar a taxa de letalidade da COVID-19 nos municípios paulistas. A plataforma de dados é composta por informações diárias referentes a taxa de letalidade, ao risco de infecção e a aplicação das vacinas como percentual da população municipal combinados com variáveis socioeconômicas como a densidade demográfica e o Índice de Desenvolvimento Humano Municipal (IDHM) para as três ondas da pandemia, sendo a primeira entre 04/02/2020 e 04/10/2020, a segunda entre 05/10/2020 e 30/11/2021, e a terceira entre 01/12/2021 e 17/03/2022.

Os principais resultados indicam que para o primeiro período de análise o IDHM possui uma relação positiva com a taxa de letalidade. Para o segundo período, um incremento de uma unidade no esquema vacinal completo (segunda dose) reduzir (mediana) em 41,67% as chances de letalidade. No terceiro período, o aumento de uma unidade na proporção de indivíduos que receberam a dose de reforço reduz as chances de letalidade em aproximadamente 78,13%.

Em geral, os resultados mostraram uma boa capacidade preditiva do modelo de regressão quantílica logística.

Por fim, esse trabalho está estruturado em quatro partes, tendo sido iniciado por essa breve introdução. Na seção 2, apresentam-se a metodologia empírica e a base de dados. Na seção 3, tem-se os resultados e na última seção são apresentadas as considerações finais.

2.METODOLOGIA E DADOS

Nessa seção apresentar-se-á o método para análise e os dados utilizados nas estimativas.

2.1 REGRESSÃO QUANTÍLICA LOGÍSTICA

Considere uma variável resposta *y*, contínua e pertencente ao intervalo $\{y_{min}, y_{max}\}$ e $x_1, \dots, x_s$ um conjunto de covariáveis. Seja $Q_y(p)$ o *p*-ésimo quantil de *y* dado o conjunto de covariáveis, em que $p \in (0,1)$. Assumindo que para qualquer quantil *p* existe um conjunto de parâmetros $\beta_p = \{\beta_{p,0}, \beta_{p,1}, \dots, \beta_{p,s}\}$ e uma função não-decrescente *h* do intervalo $[y_{min}, y_{max}]$ para a reta real, pode-se escrever o modelo da seguinte forma:

$$h\{Q_y(p)\} = \beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s. \quad (1)$$

Considerando uma variável resposta limitada no intervalo unitário, o que se assemelha a uma probabilidade em vários aspectos, tem-se que uma escolha razoável para a função *h* é o *link* logístico (BOTTAI *et al*., 2010):

$$h(y) = logit(y) = \log\left(\frac{y - y_{min}}{y_{max} - y}\right). \quad (2)$$

Consequentemente:

$$Q_y(p) = \frac{\exp(\beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s)y_{max} + y_{min}}{\exp(\beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s) + 1}. \quad (3)$$

O modelo regressivo quantílico desempenha um papel importante, sendo capaz de modelar quantis condicionais em função de variáveis explicativas. Além disso, a abordagem quantílica considerada é vantajosa uma vez que não há é necessário assumir uma distribuição de probabilidade para os dados (DAVINO et al. 2022).

Na regressão linear, as estimativas são obtidas por meio da minimização da função do erro quadrático médio. Como na regressão quantílica o foco é no *p-ésimo* quantil, uma outra função de perda precisa ser utilizada. Conforme, Koenker e Machado (1999), a função de perda deve ser:

$$l_p(u) = \{up, se\ u \geq 0; u(p-1), se\ u < 0\}. \quad (4)$$

Logo, para obter-se o conjunto de estimativas dos parâmetros, resolve-se o seguinte problema de minimização:

$$min_{\beta \in R^{q+1}} \sum_{i=1}^{n} l_p[y_i - (\beta_{p,0} + \beta_{p,1}x_1 + \cdots + \beta_{p,s}x_s)]. \quad (5)$$

O método utilizado para minimização de (5) é uma versão modificada do algoritmo de Barrodale e Roberts para regressão L1; ver detalhes em Koenker e d'Orey (1987, 1994).

2.2 DADOS

O banco de dados foi constituído por três fontes principais. A primeira refere-se às informações relativas ao número de casos de COVID-19, as quais foram fornecidas pela Secretaria de Saúde do Estado de São Paulo (SES-SP). Os microdados fornecidos pela SES-SP contém informações sobre os fatores de risco aos quais se deparava o paciente e se o indivíduo faleceu.

Adicionalmente, utilizou-se informações referentes a vacinação dos indivíduos disponibilizadas no portal do Ministério da Saúde, OpenDataSUS. A partir dessas informações é possível calcular o número de doses aplicadas em cada indivíduo e o número (1º, 2ª ou 3ª dose). As informações socioeconômicas, como população, tamanho do território do município foram obtidas a partir do Instituto Brasileiro de Geografia e Estatística (IBGE).

Como a situação epidemiológica é sensível ao período analisado, dividiu-se a amostra em três períodos distintos: i) o primeiro período está situado entre 04/02/2020 a 04/10/2020; ii) o segundo concentra-se entre 05/10/2020 a 30/11/2021, e por fim iii) o terceiro de 01/12/2021 a 17/03/2022. Essa divisão teve como objetivo capturar os diferentes períodos de aplicações das doses vacinais. A Tabela 1, faz um resumo das variáveis utilizadas.

Tabela 1 – Resumo das variáveis utilizadas

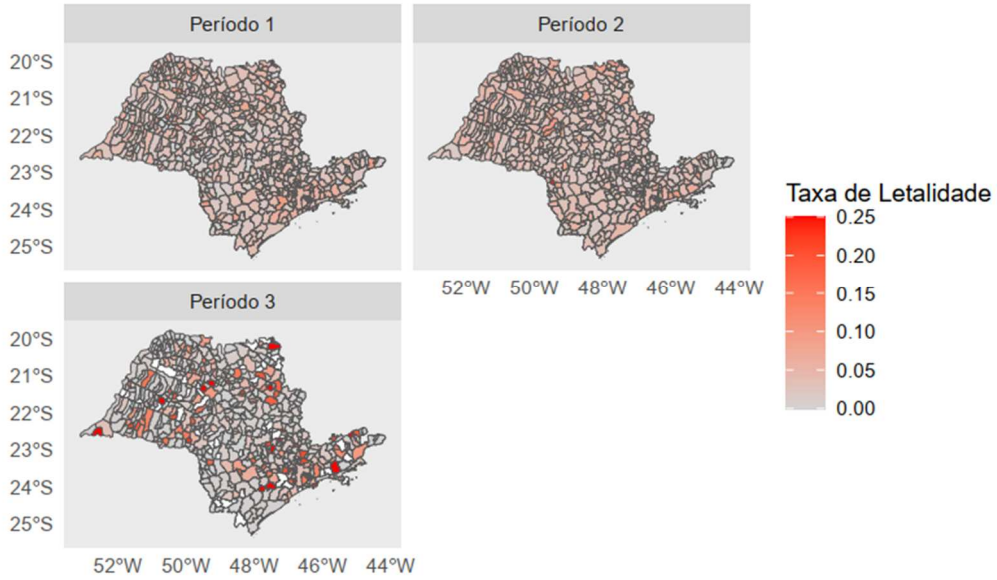
Variável	Definição	Fonte
Taxa de Letalidade	Número de mortes dividido pelo número total de casos	SES-SP
Risco	A porcentagem de casos de covid-19 com alguma complicação de risco (asma, diabetes, cardiopatia, doença hematológica, hepática,	SES-SP

	neurológica entre outros).	
Idade Mediana	Idade mediana (em anos) dos pacientes com COVID-19 no município durante cada período.	SES-SP
Densidade Demográfica 2021	Densidade Populacional do município (em habitantes/km^2)	IBGE
Índice de Desenvolvimento Humano	Fornecido pelo IBGE em 2020. É um índice que varia entre 0 e 1.	IBGE
Dose 1	Porcentagem normalizada pelo método min-max da população municipal com pelo menos uma dose da campanha de vacinação, sem contabilizar as doses únicas.	OpenDataSUS e IBGE
Dose 2	Porcentagem normalizada pelo método min-max da população municipal totalmente vacinada de acordo com o antigo plano vacinal dentro de cada intervalo temporal	OpenDataSUS e IBGE
Reforço	Porcentagem normalizada pelo método min-max da população local vacinada com a primeira dose de reforço contra a COVID-19.	OpenDataSUS e IBGE.

Fonte - Elaborado pelos autores

A Figura 1 mostra a evolução da taxa de letalidade da COVID-19 no período analisado.

Figura 1 – Taxa de Letalidade para o Estado de São Paulo



Fonte - Elaborado pelos autores

3. RESULTADOS

Nessa seção será realizada a análise exploratória dos dados, bem como, as estimações para os três períodos estudados. Todos os códigos elaborados para o tratamento dos dados e para modelagem estão disponíveis em R (https://www.r-project.org/) no seguinte endereço:

https://github.com/Torm198/RegressaoQuantilicaCovid.

Particularmente, os pacotes do R utilizados no tratamento foram: tidyverse, readr, ggcorrplot, geobr, gtools, e lubridate; já os pacotes do R para ajuste dos modelos foram: quantreg, rms e VGAM.

3.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Inicialmente far-se-á a análise da variável taxa de letalidade (let). Na Tabela 2, apresentam-se as estatísticas descritivas desse indicador, dentre elas são encontrados os coeficientes de variação (CV), assimetria (CA) e excesso de curtose (CC). Ressalta-se que o tamanho da amostra no período 3 é menor devido a alguns municípios não terem apresentado casos.

Tabela 2 – Estatísticas sumárias para as taxas de letalidade

Período	Média	Mediana	DP	CV	CA	CC	min	max	N
Período 1	0,030	0,027	0,022	73,23	2,05	12,14	0,00	0,20	645
Período 2	0,032	0,029	0,018	56,50	3,45	31,62	0,002	0,24	645
Período 3	0,070	0,010	0,138	196,40	2,92	12,58	0,00	1,00	636

Fonte - Elaborado pelos autores

Da Tabela 2, observa-se que a taxa de letalidade apresenta assimetria positiva e alto grau de curtose. Observa-se ainda que o último período possui uma menor mediana e um CV de 196,4%, ou seja, possui uma maior dispersão dos dados. É importante destacar que o número de pacientes infectados com a COVID-19 diminuiu no estado no terceiro período, o que permite uma maior variabilidade nos dados.

3.2 ANÁLISE DAS ESTIMATIVAS

Nessa seção, os dados da taxa de letalidade da COVID-19 dos municípios paulistas são modelados por meio do modelo de regressão logística para dados limitados de Bottai et al. (2010).

3.2.1 ESTIMATIVAS PARA O PERÍODO 1

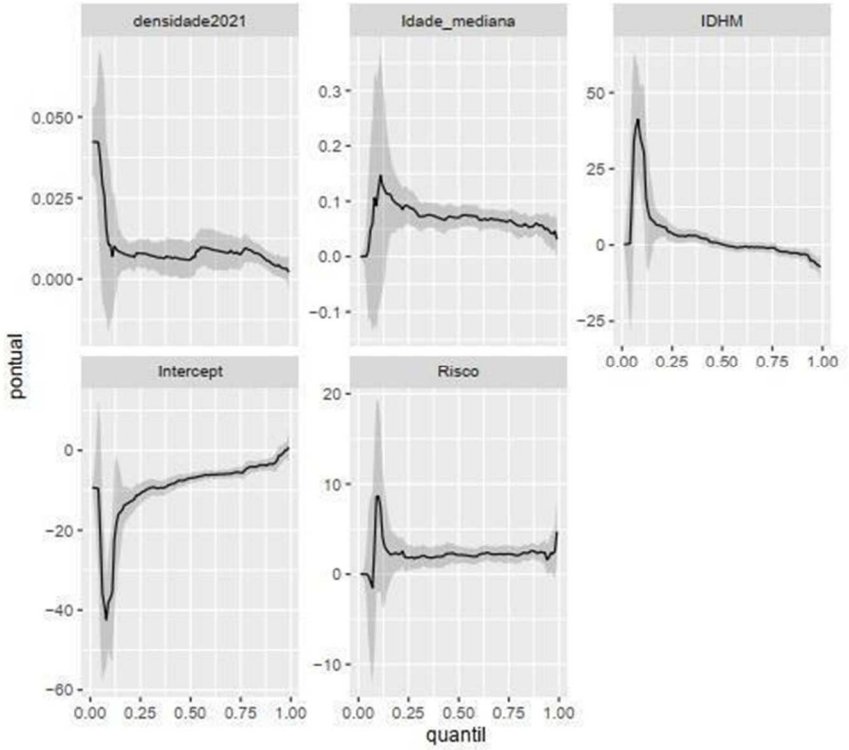
Como o primeiro período precede o início da campanha nacional de vacinação não foram incluídas as variáveis referentes à aplicação da primeira e da segunda doses nessas regressões. Uma análise do fator de inflação de variância (VIF) para cada variável explicativa indica ausência de multicolinearidade (nenhum VIF apresentou um valor acima de 10). A Tabela 3 apresenta as estimativas dos parâmetros para alguns quantis, e a Figura 2 apresenta as estimativas dos parâmetros do modelo de regressão quantílica e os intervalos de confiança de 95%, obtido por meio da estimação do erro padrão pelo método de *bootstrap*. Observa-se que há dinâmicas assimétricas com estimativas mudando em q e algumas estimativas mudam de sinal.

Tabela 3 – Estimativas (erro padrão entre parênteses) para alguns quantis para o Período 1

Quantil	0,05	0,1	0,25	0,5	0,75	0,9	0,95
Intercepto	-21,181* (0,021)	-37,028* (0,028)	-10,8* (0,016)	-6,935* (0,016)	-5,58 (0,022)	-3,368 (0,033)	-1,208* (0,054)
IDHM	15,982* (0,027)	32,95* (0,038)	3,81* (0,021)	0,246 (0,02)	-0,928* (0,026)	-3,035* (0,04)	-5,107* (0,057)
densidade2021	0,037* (<0,001)	0,011* (<0,001)	0,008* (<0,001)	0,006** (<0,001)	0,008* (<0,001)	0,005 (<0,001)	0,004* (<0,001)
Risco	-0,169 (0,005)	8,636 (0,021)	1,811* (0,011)	2,128* (0,012)	2,264* (0,017)	2,376* (0,026)	1,859* (0,032)
Idade_mediana	0,013 (<0,001)	0,122 (<0,001)	0,091* (<0,001)	0,07* (<0,001)	0,065* (<0,001)	0,056* (0,001)	0,047* (0,001)

Fonte - Elaborado pelos autores
* significante a 5%. ** significante a 10%.

Figura 2 – Estimativas dos coeficientes com intervalo de confiança de 95% ao longo dos quantis para o Período 1



Fonte - Elaborado pelos autores

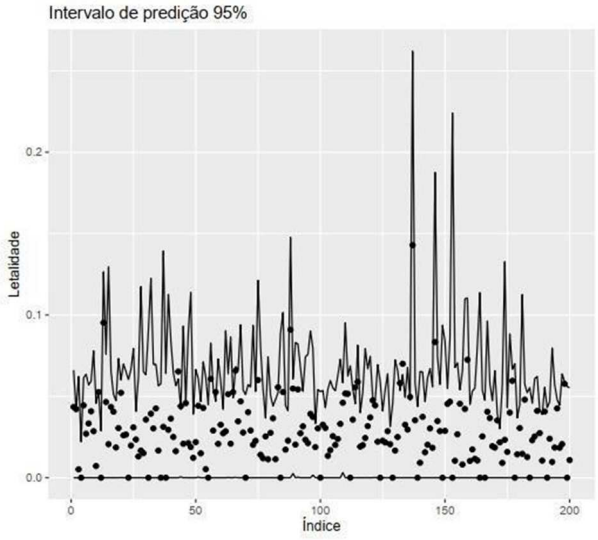
Os resultados da Figura 2 revelam que, em parte do espectro quantílico, as covariáveis impactam positivamente a taxa de letalidade. Em particular, o IDHM apresenta o efeito mais elevado para os quantis mais baixos, o que é decorrente, provavelmente, do fato que municípios mais desenvolvidos receberam mais pacientes infectados pela COVID-19, o que acarreta numa taxa de letalidade mínima mais alta. Por outro lado, há a possibilidade que esse efeito positivo ser decorrente da variância do coeficiente na região. Ao observarem-se os quantis mais altos essa variável começa a ter um efeito negativo sobre a taxa de letalidade indicando que municípios com o IDHM mais elevado possuem uma capacidade mais adequada para lidarem com os casos de COVID-19.

O coeficiente IDHM para o quantil *q=0,50* (mediana) apresentou o valor estimado de aproximadamente 0,246. Em termos de razão de chance, esse coeficiente pode ser interpretado da seguinte forma: (exp(0,246)-1)x100% = 27,89%. Isto é, para cada incremento de uma unidade no IDHM aumenta-se em 27,89% as chances de letalidade da COVID-19.

O coeficiente da variável risco assumiu um valor negativo por um breve momento nos quantis inferiores, seguido por um pico positivo, e manteve-se positivo ao longo dos quantis restantes. Nas demais variáveis explicativas, observou-se uma relação positiva, ou seja, quanto maior a idade mediana e a densidade populacional maior é a taxa de letalidade.

A Figura 3 apresenta o intervalo de predição de 95% do modelo para o período 1. A construção dessa forma preditiva é feita da seguinte forma: para cada valor de $q \in \{0.025, 0.975\}$, os parâmetros do modelo foram estimados e as respectivas estimativas foram utilizadas para obter as predições para 0.025 e 0.975. As predições foram realizadas 200-passos-à-frente, ou seja, 200 observações são retiradas aleatoriamente do banco e não foram incluídas na estimação. A Figura 3 mostra que 89.5% das observações estão dentro dos limites do intervalo de predição.

Figura 3 – Intervalos de predição de 95% para o Período 1



Fonte - Elaborado pelos autores

3.2.2 ESTIMATIVAS PARA O PERÍODO 2

No segundo período a campanha de vacinação já estava ocorrendo e o esquema vacinal completo era considerado quando o indivíduo havia recebido a primeira e a segunda dose da vacina. Por esse motivo, apenas foi incluída a variável que identifica a segunda dose no modelo estimado. A Tabela 4 apresenta as estimativas dos parâmetros para alguns quantis, e a Figura 4 apresenta as estimativas dos coeficientes além do intervalo de confiança de 95%.

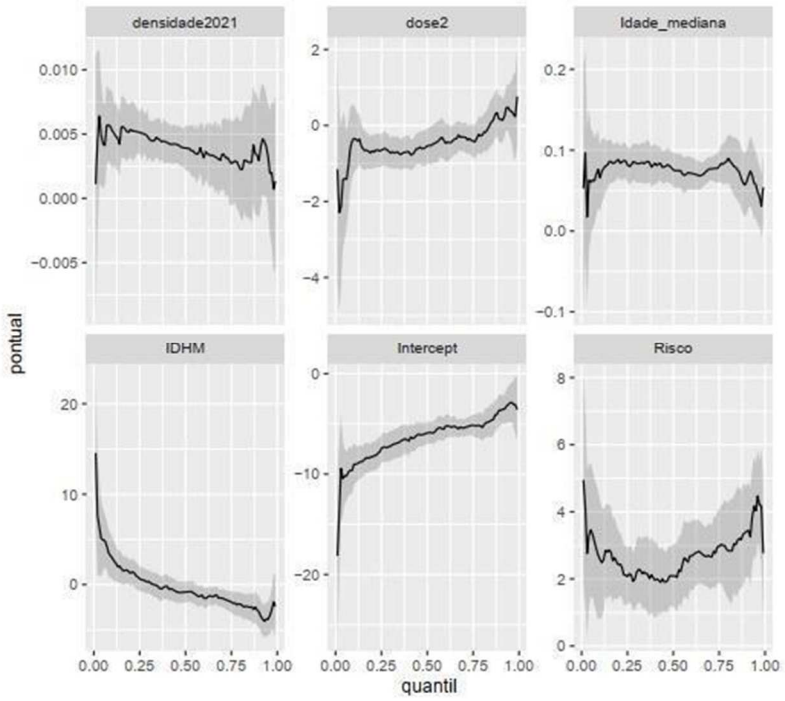
Tabela 4 – Estimativas (erro padrão entre parênteses) para alguns quantis para o Período 2

Quantil	0,05	0,1	0,25	0,5	0,75	0,9	0,95
Intercept	-10,169* (0,027)	-9,021* (0,018)	-7,45* (0,011)	-5,894* (0,015)	-5,151 (0,013)	-3,651 (0,034)	-2,925 (0,037)
IDHM	4,938* (0,03)	2,972* (0,021)	0,644 (0,013)	-0,827* (0,016)	-1,939* (0,017)	-3,007* (0,035)	-3,832* (0,044)
densidade2021	0,004 (<0,001)	0,005* (<0,001)	0,005* (<0,001)	0,004* (<0,001)	0,003* (<0,001)	0,003 (<0,001)	0,003 (<0,001)
Risco	3,462* (0,019)	2,583* (0,015)	2,072* (0,009)	2,088* (0,015)	2,97* (0,018)	3,42* (0,032)	4,023* (0,036)

Idade_mediana	0,06** (<0,001)	0,066* (<0,001)	0,082* (<0,001)	0,076* (<0,001)	0,08* (<0,001)	0,061* (0,001)	0,057* (0,001)
dose2	-1,395 (0,01)	-0,342 (0,007)	-0,65* (0,003)	-0,539* (0,006)	-0,412 (0,007)	0,164 (0,014)	0,388** (0,014)

Fonte - Elaborado pelos autores
* significante a 5%. ** significante a 10%.

Figura 4 – Estimativas dos coeficientes com intervalo de confiança de 95% ao longo dos quantis para o Período 2



Fonte - Elaborado pelos autores

Como viu-se nas estimativas para o primeiro período, o intercepto e os coeficientes atrelados aos IDHM e a idade mediana apresentaram um comportamento similar. Em relação ao coeficiente da densidade populacional, este apresentou uma característica negativa na extrema direita do espectro, o que pode ser explicado tanto pela variância, indicando um baixo poder preditivo nessa região, quanto pelo fato das regiões mais densas no estado não apresentarem uma taxa de letalidade mais alta, assim tendo um limite superior menor.

A variável que se refere à segunda dose interage de forma negativa em relação à taxa de letalidade em boa parte do espectro temporal avaliado. O comportamento positivo da taxa de letalidade pode ser explicado pela

possibilidade dos casos e óbitos em locais onde há alta taxa de mortalidade de pacientes não vacinados em áreas com taxa de aplicação de vacina alta, porém sem ainda completar o esquema vacinal de uma parcela da população. Assim, grande parte das notificações dos casos de COVID-19 podem ser de um grupo mais sensível em relação àquele totalmente imunizado.

Em termos da razão de chances, o aumento de uma unidade no percentual de cidadãos que tomaram a segunda dose no quantil *q=0,50* (mediana), reduz em 41,67% as chances de letalidade.

Assim como no primeiro período realizou-se uma análise da performance de predição para um intervalo de 95%. Foram sorteadas 200 observações do banco para compor a amostra de validação (predição 200 passos-à-frente). Aqui, 94,5% das observações estão dentro do intervalo, valor muito próximo ao valor nominal de 95%.

3.2.3 ESTIMATIVAS PARA O PERÍODO 3

No terceiro período, foram realizadas estimativas com a variável reforço, isto é, todos os indivíduos que tomaram a segunda dose e receberam uma vacina adicional. A Tabela 5 apresenta as estimativas dos parâmetros para alguns quantis. A Figura 5 apresenta as estimativas dos parâmetros do modelo de regressão quantílica e os intervalos de confiança de 95%. Algumas variáveis apresentaram um comportamento esperado como a variável de reforço, que tem um efeito negativo em relação a taxa de letalidade e as variáveis risco e idade mediana que apresentam um efeito positivo. O IDHM e a densidade populacional apresentaram um efeito positivo, mas decrescentes ao longo dos quantis.

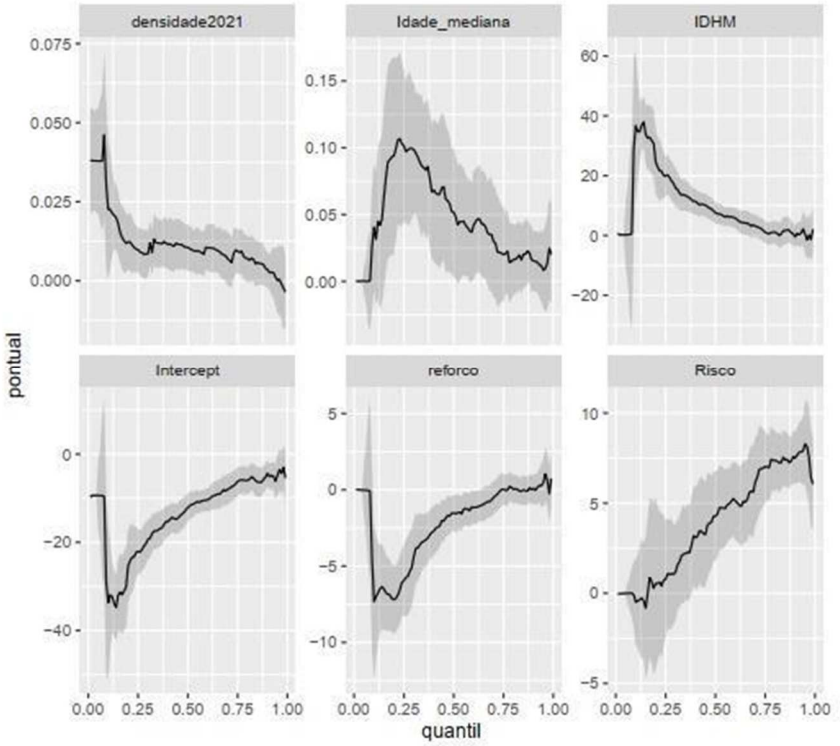
Tabela 5 – Estimativas (erro padrão entre parênteses) para alguns quantis para o Período 3

Quantil	0,05	0,1	0,25	0,5	0,75	0,9	0,95
Intercept	-9,392* (0)	-33,739 (0,01)	-22,085* (0,019)	-12,1* (0,009)	-6,218 (0,027)	-4,355 (0,056)	-4,767 (0,07)
IDHM	0,233* (0)	36,737 (0,015)	19,852* (0,016)	7,586 (0,012)	1,072* (0,034)	-0,483* (0,068)	0,507 (0,089)
densidade2021	0,038* (<0,001)	0,023 (<0,001)	0,01* (<0,001)	0,011 (<0,001)	0,009 (<0,001)	0,004 (0,001)	<0,001* (<0,001)
Risco	-0,006 (<0,001)	-0,491 (<0,001)	0,779* (0,029)	4,274* (0,023)	7,033* (0,041)	7,503* (0,086)	8,287* (0,063)

Idade_mediana	<0,001 (<0,001)	0,041 (<0,001)	0,101* (<0,001)	0,052* (<0,001)	0,02* (<0,001)	0,015* (0,001)	0,008 (0,001)
reforco	-0,018 (<0,001)	-7,33** (0,002)	-5,936* (0,004)	-1,52* (0,002)	0,036 (0,008)	0,098 (0,015)	0,454 (0,022)

Fonte **-** Elaborado pelos autores
* significante a 5%. ** significante a 10%.

Figura 5 – Estimativas dos coeficientes com intervalo de confiança de 95% ao longo dos quantis para o Período 3



Fonte - Elaborado pelos autores

Baseado nos resultados da Tabela 5, nota-se que o aumento de uma unidade na variável reforço no quantil *q=0.50* (mediana), está associado a uma diminuição de 78,13% nas chances de letalidade.

Para um intervalo de predição de 95% do modelo para o período 3, com predições realizadas 200-passos-à-frente, os resultados mostraram que 97,5% das observações estão dentro dos limites do intervalo de predição, o que é bem próximo ao valor nominal de 95%.

4. CONSIDERAÇÕES FINAIS

A pandemia do COVID-19 foi responsável por uma das maiores crises sanitárias e econômica que a humanidade enfrentou nos últimos séculos. Dentro deste contexto, esse estudo utilizou um modelo de regressão quantílica logística para a taxa letalidade da COVID-19 nos municípios de São Paulo. Para isso montou-se uma plataforma de dados que combina informações diárias sobre o risco de infecção e a aplicação das vacinas como percentual da população municipal com dados socioeconômicas como a densidade demográfica e o Índice de Desenvolvimento Humano Municipal (IDHM) para as três ondas da pandemia ocorridas entre 2020 e 2022.

Os principais resultados indicam que os municípios mais desenvolvidos, com IDHM mais alto, em média, possuíam uma maior taxa de letalidade na primeira onda. Em relação à segunda e a terceira onda, as estimativas indicaram que a razão de vacinados pelo total da população municipal apresentou um relacionamento negativo com a taxa de letalidade, o que pode ser um indício do sucesso da aplicação do esquema vacinal para o enfrentamento da pandemia. Além disso, os resultados mostraram uma boa capacidade preditiva do modelo de regressão quantílica logística.

Em suma, para uma nova agenda de pesquisa ainda há espaço para análises longitudinais, o que traria mais riqueza de resultados, além da expansão da análise a outros estados brasileiros.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alvarez, F. E., Argente, D. & Lippi, F. (2020). A Simple Planning Problem for COVID-19 Lockdown. Working Paper, National Bureau of Economic Research.

Bottai, M., Cai, B. & Mckeown R. E. (2010). Logistic quantile regression for bounded outcomes. Statistics in Medicine, 29(2):309–317.

Brown, S. (2008). Global Public-Private Partnerships for Pharmaceuticals: Operational and Normative Features, Challenges, and Prospects. Canadian Public Administration Journal.

Dave, D.M., Friedson, A.I. & Matsuzawa, K.; YOU KNOW, J.J. (2020). When Do Shelter-in-Place Orders Fight COVID-19 Best? Policy Heterogeneity Across States and Adoption Time. Working Paper, National Bureau of Economic Research.

Davino, C., Furno, M. & Vistocco, D. (2014). Quantile Regression. Wiley, Chichester.

Davino, C., Romano, R. & Vistocco, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. METRON.

Eicheinbaum, M. S., Rebelo, S. & Trabandt, M. (2020). The Macroeconomics of Epidemics. Working Paper, National Bureau of Economic Research.

Farboodi, M., Jarosch, G. & Shimer, R. (2020). Internal and External Effects of Social Distancing in a Pandemic. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.

Ferrari, S. & Cribari-Neto. (2014). Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31:799–815.

Fong, M. W., Gao, H., Wong, J. Y., Xiao, J., Shiu, E. Y. C, Ryu, S. & Cowling, B. J. (2020). Nonpharmaceutical Measures for Pandemic Influenza in Nonhealthcare Settings—Social Distancing Measures. Emerging Infectious Diseases Journal, v. 26, n. 5.

Friedson, A., Mcnichols, D., Sabia, J. & Dave, D. (2020). Did California's Shelter-in-Place Order Work? Early Coronavirus-Related Public Health Effects. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.

Gupta, S., Nguyen, T. D., Rojas, F. L.; et al. (2020). Tracking Public and Private Responses to the COVID-19 Epidemic: Evidence from State and Local Government Actions. Working Paper, National Bureau of Economic Research.

Oliveira, C. (2020). Does "Staying at Home" Save Lives? An Estimation of the Impacts of Social Isolation in the Registered Cases and Deaths by COVID-19 in Brazil. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.

Oliveira, C, Pereira, R. M. & Costeira, G. M. (2020). Avaliando os custos e benefícios da intensificação do isolamento social no Rio Grande do Sul a partir de um experimento natural. Available on: < https://www.researchgate.net/publication/342666325_Avaliando_os_custos_e_beneficios_da_intensificacao_do_isolamento_social_no_Rio_Grande_do_Sul_a_partir_de_um_experimento_natural >. Accessed on: 17/07/2020.

Koenker, R. & Bassert JR, G. (1978). Regression quantiles. Econometrica, 46:33–50.

Koenker, R. & D'Orey, V. (1987). Computing regression quantiles. Applied Statistics, 36:383– 393.

Koenker, R & D'Orey, V. (1984). Computing regression quantiles. Applied Statistics, 43:410–414.

Koenker, R. & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association, 94(448):1296–1310.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

SEADE. População paulista cresceu 20% em 20 anos (2022). https://www.seade.gov.br/populacao-paulista-cresceu-20-em-20-anos/, acessado em 22/04/2022

AGRADECIMENTOS e COLABORAÇÕES

Os autores agradecem o financiamento do CNPq, CAPES e FAP-DF.

ISSN 2675-3243

volume 80

número 247

Janeiro/dezembro 2022