

Um estudo sobre a evolução de óbitos por câncer de mama no Brasil usando modelos de séries temporais

A study on the evolution of breast cancer deaths in Brazil using time series models

Un estudio sobre la evolución de las muertes por cáncer de mama en Brasil utilizando modelos de series temporales

Recebido: 23/12/2020 | Revisado: 25/12/2020 | Aceito: 28/12/2020 | Publicado: 30/12/2020

Rafael Agostinho Ferreira

ORCID: <https://orcid.org/0000-0003-0998-8767>

Universidade Federal de Lavras, Brasil

E-mail: rafaelferreiratc@hotmail.com

Vânia de Fátima Lemes de Miranda

ORCID: <https://orcid.org/0000-0002-0624-6840>

Universidade Federal de Lavras, Brasil

E-mail: vaniafamat@gmail.com

Patrícia Mendes dos Santos

ORCID: <https://orcid.org/0000-0001-6158-1724>

Universidade Federal de Lavras, Brasil

E-mail: patymendesdossantos@hotmail.com

Henrique José de Paula Alves

ORCID: <https://orcid.org/0000-0002-0124-3093>

Instituto de Pesquisa Econômica Aplicada, Brasil

E-mail: jpahenrique@gmail.com

Thelma Sáfadi

ORCID: <https://orcid.org/0000-0002-4918-300X>

Universidade Federal de Lavras, Brasil

E-mail: safadi@ufla.br

Resumo

O câncer de mama é o tipo de câncer mais comum entre as mulheres e é a principal causa de morte no mundo, está entre os cinco cânceres mais incidentes no Brasil. Diante disso, entende-se que é de grande importância a avaliação do número de óbitos no Brasil, uma vez que conhecer o comportamento da doença é fundamental para os órgãos públicos voltados

para a saúde e bem estar da população. Desse modo, este trabalho tem como objetivo utilizar técnicas de séries temporais para analisar a série de observações referentes ao número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2018. Para esta análise, foram verificadas a variabilidade da série, assim como a presença das componentes tendência e sazonalidade. A metodologia Box-Jenkins foi utilizada na modelagem dos dados, sendo que a série em estudo foi bem ajustada utilizando modelos da classe SARIMA. A comparação entre os modelos considerados adequados à série foi realizada através do AIC e EQMP. O modelo ajustado foi utilizado para fazer previsões sobre as observações futuras dessa série. De acordo com essa previsão, foi possível observar que para os meses subsequentes a série manterá o padrão que ela vem mantendo desde o início do seu período de observação: um crescente aumento no número de óbitos por tal enfermidade.

Palavras-chave: Óbitos; Modelo SARIMA; Box-Jenkins; Previsão.

Abstract

Breast cancer is the most common type of cancer among women and is the leading cause of death worldwide. It is among the five most incident cancers in Brazil. Given this, it is understood that it is important to assess the number of deaths in Brazil since knowing the behavior of the disease is essential for public agencies focused on the health and well-being of the population. Thus, this study aims to use time series techniques to analyze the number of observations regarding the number of deaths from breast cancer (group ICD-10: Malignant neoplasms in the breast) in Brazil, covering the period from January 1996 to December 2018. For this analysis, the variability of the series was verified and the presence of the trend and seasonality components. The Box-Jenkins methodology was used to model the data, and the series under study was well adjusted using models of the SARIMA class. The comparison between the models considered suitable for the series was performed using the AIC and EQMP. The adjusted model was used to make predictions about future observations in this series. According to this forecast, it was possible to observe that for the following months, the series will maintain the pattern it has been maintaining since the beginning of its observation period: a growing increase in the number of deaths from such disease.

Keywords: Deaths; SARIMA model; Box-Jenkins; Forecast.

Resumen

El cáncer de mama es el tipo de cáncer más común entre las mujeres y es la principal causa de muerte en todo el mundo. Se encuentra entre los cinco cánceres más incidentes en Brasil. Ante esto, se entiende que es importante evaluar el número de muertes en Brasil ya que conocer el comportamiento de la enfermedad es fundamental para las agencias públicas enfocadas en la salud y el bienestar de la población. Así, este estudio tiene como objetivo utilizar técnicas de series temporales para analizar el número de observaciones sobre el número de muertes por cáncer de mama (grupo CIE-10: Neoplasias malignas en la mama) en Brasil, cubriendo el período de enero de 1996 a diciembre de 2018. En este análisis se verificó la variabilidad de la serie y la presencia de los componentes tendencial y estacionalidad. Se utilizó la metodología Box-Jenkins para modelar los datos y la serie en estudio se ajustó bien utilizando modelos de la clase SARIMA. La comparación entre los modelos considerados adecuados para la serie se realizó mediante el AIC y el EQMP. El modelo ajustado se utilizó para hacer predicciones sobre observaciones futuras en esta serie. De acuerdo con esta previsión, se pudo observar que para los próximos meses la serie mantendrá el patrón que viene manteniendo desde el inicio de su período de observación: un incremento creciente en el número de muertes por dicha enfermedad.

Palabras clave: Muertes; Modelo SARIMA; Box-Jenkins; Pronóstico.

1. Introdução

Uma série temporal é um conjunto de observações ordenadas no tempo. Por serem dados ordenados no tempo, espera-se que as observações próximas sejam dependentes e, por essa razão, resultam em análises cujos erros apresentam uma estrutura de dependência temporal. Isso impossibilita a aplicação de métodos usuais que dependem da suposição de que as observações próximas sejam independentes e identicamente distribuídas.

Dessa forma, o objetivo da modelagem de séries temporais é analisar e modelar a correlação existente entre os erros. Assim, a dependência entre os erros deixa de ser vista como inapropriada, uma vez que esta informação é incorporada ao modelo. Para modelar uma série temporal, a metodologia mais utilizada foi a proposta por Box & Jenkins (1976), que se baseia na suposição de que uma observação, e/ou seu erro, em um determinado momento é influenciada pelo que ocorreu em um passado recente. Esses modelos são conhecidos, na literatura, como modelos Seasonal Autoregressive Integrated Moving Average (SARIMA).

No presente trabalho, a análise de séries temporais foi utilizada para descrever as observações do número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2018.

O câncer de mama é o tipo de câncer mais comum entre as mulheres e é a principal causa de morte no mundo, está entre os cinco cânceres mais incidentes no Brasil.

Entre várias outras doenças, o câncer é uma das que mais provoca impacto psicológico, já que muitas vezes representa ou é vista como uma caminhada progressiva e dolorosa em direção à morte e/ou à mutilação (Silva, 2005). Dentre os fatores de risco para o desenvolvimento do câncer de mama, estão os fatores reprodutivos, a história de ciclos menstruais, a multiparidade, o uso de terapia hormonal pós-menopausa ou contraceptivos orais e a paridade tardia. O consumo de álcool e fatores dietéticos representam cerca de 30% das causas de câncer, sendo superado somente pelo tabaco. E ainda, a gênese e progressão do câncer de mama parecem estar extremamente relacionadas a hábitos alimentares, consumo de gorduras, carnes, produtos lácteos, frutas, vegetais, fibras, fitoestrógenos e outros componentes dietéticos (Padilha; Pinheiro, 2004).

Barbosa et al. (2015), constatou que o câncer também pode estar relacionado ao status socioeconômico: mulheres com mais alto status socioeconômico possuem maior risco de desenvolver câncer de mama que mulheres do grupo com status socioeconômico abaixo da média.

A mamografia foi estabelecida no Brasil em abril de 2004, para mulheres entre 50 e 69 anos de idade, com o intervalo máximo de dois anos entre os exames Barbosa et al. (2015). Os programas para detecção precoce do câncer de mama, em sua maioria apoiam-se na tríade constituída pelo autoexame das mamas, exame clínico das mamas e mamografia (Batiston et al., 2009).

Por outro lado, há estudos que mostram não identificar a redução na taxa de mortalidade com a aplicação da mamografia como método de detecção da doença, antes mesmo da pessoa apresentar sintomas, onde seria permitido o tratamento em fases iniciais.

Nesse sentido, entende-se que a utilização de técnicas de séries temporais é de grande importância para a avaliação do número de óbitos no Brasil, uma vez que conhecer o comportamento da doença é fundamental para os órgãos públicos voltados para a saúde e bem estar da população. Dessa forma, após descrever o comportamento do número de óbitos por meio do ajuste de um modelo de séries temporais, o mesmo foi utilizado para fazer previsões para os óbitos referentes aos anos de 2019 e 2020.

2. Material e Métodos

A série em estudo consiste em observações referentes ao número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2018. Os dados foram obtidos por meio de consulta às bases de dados do Departamento de Informática do Sistema Único de Saúde (DATASUS) no endereço eletrônico <http://www.data-sus.gov.br>. Deste modo, este estudo trata-se de uma pesquisa exploratória, de natureza quantitativa, uma vez que técnicas estatísticas foram utilizadas para análise e modelagem de dados, como descrito por Pereira, Shitsuka, Parreira e Shitsuka (2018).

Uma série temporal é descrita como um processo estocástico (Morettin; Toloi, 2006). Ou seja, é uma coleção de variáveis aleatórias indexadas à um conjunto qualquer (Ross, 1996). Este conjunto de indexação pode ser discreto ou contínuo. Geralmente, para uma série temporal, este conjunto de indexação é discreto e é definido como o próprio tempo. Formalmente, define-se a série temporal \mathfrak{S} como

$$\mathfrak{S} = Z_t, t \in T \subset \mathbb{Z}$$

em que Z_t é uma variável aleatória indexada no tempo t , o qual pertence ao conjunto de indexação T que está contido no conjunto dos inteiros \mathbb{Z} .

A realização deste processo estocástico gera um conjunto de observações sequenciadas no tempo Z_t , resultante de um modelo. Morettin e Toloi (2006) apresenta uma proposta de modelagem para Z_t o qual a descreve pela combinação de três componentes,

$$Z_t = T_t + S_t + a_t \quad (1)$$

em que T_t representa uma tendência, S_t uma sazonalidade e a_t é uma componente aleatória, com média 0 centrada em e variância constante σ^2 .

Para ajustar os modelos da série temporal, inicialmente, construiu-se o gráfico da série original para verificar a existência de possíveis componentes (tendência e ou sazonalidade) ou observações atípicas. Entretanto, antes de inferir sobre as componentes da série, foi necessário verificar se o modelo é aditivo ou multiplicativo e se apresenta variância não constante. Em caso afirmativo, uma transformação Box-Cox se faz necessária a fim de estabilizar a variância e tornar o modelo aditivo (Morettin; Toloi, 2006). A transformação Box-Cox é dada por:

$$Z_t^*(\lambda) = \begin{cases} \frac{Z_t^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0 \\ \log Z_t, & \text{se } \lambda = 0 \end{cases} \quad (2)$$

Para analisar a existência de tendência na série investigou-se o gráfico da função de autocorrelação da série, cujo decaimento lento para zero indica indícios da presença de tendência. Para confirmar a significância dessa componente utilizou-se teste de Cox-Stuart (ou Teste dos Sinais) proposto por Cox e Stuart (1955).

Para verificar a existência da sazonalidade, foram testadas as seguintes hipóteses:

H_0 : não existe sazonalidade;

H_1 : existe sazonalidade.

Foi utilizado o teste não-paramétrico de Friedman, proposto, à princípio, para testar diferenças entre blocos em análises de experimentos. Este teste consiste em considerar os meses como tratamentos e os anos como blocos (Morettin; Toloí, 2006). A estatística de teste é dada por:

$$T_c = \frac{12}{pk(k+1)} \sum_{j=1}^k R_{.j}^{*2} - 3p(k+1) \quad (3)$$

em que p é o número de anos considerados, k é o número de meses e $R_{.j}^*$ é a soma dos postos da j -ésima coluna. Mais detalhes podem ser encontrados em Morettin e Toloí (2006).

Uma vez identificada a presença das componentes de tendência e/ou sazonalidade, pode-se retirá-la(s) pelo método da diferenciação. Para a remoção de tendência, aplica-se a diferenciação de ordem. Do mesmo modo, para a remoção da componente sazonal, aplica-se uma diferenciação de ordem na série. Uma vez removidas essas componentes, obtêm-se uma série estacionária. A diferenciação é geralmente usada para simplificar a estrutura de correlação ou revelar padrões, preparando as séries com tendência e sazonalidade para ajuste de modelos ARIMA (Morettin; Toloí, 2006).

No entanto, a série resultante, após as diferenciações, pode apresentar uma estrutura de correlação temporal significativa, seja de baixa ordem ou em ordens de período sazonal (Morettin; Toloí, 2006). Neste caso, se faz necessário o ajuste de um modelo sob a série original, que considere uma sazonalidade estocástica. Um modelo proposto por Box e Jenkins

(1976), o qual considera este comportamento, é o modelo ARIMA sazonal (ou SARIMA). Um modelo SARIMA(p,d,q)(P,D,Q)_s é descrito como

$$\Phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^d Z_t = \Theta(B^s)\theta(B)e_t \quad (4)$$

em que

• $\Phi(B^s) = 1 - \phi_1 B^{s1} - \phi_2 B^{s2} - \dots - \phi_p B^{sP}$ é o operador autorregressivo sazonal de ordem p , estacionário;

• $\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \dots - \phi_p B^p$ é o operador autorregressivo não sazonal de ordem p , estacionário;

• $\Theta(B^s) = 1 - \theta_1 B^{s1} - \theta_2 B^{s2} - \dots - \theta_Q B^{sQ}$ é o operador de médias móveis sazonal de ordem Q , invertível;

• $\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \dots - \theta_q B^q$ é o operador de médias móveis não sazonal de ordem q , invertível;

• $(1 - B^s)^D$ é o operador de diferenças sazonais de ordem D ;

• $(1 - B)^d$ é o operador de diferenças não sazonais de ordem d ;

• e_t é um ruído branco.

O processo de identificação da ordem do modelo SARIMA, apresentado na equação (4), é feito baseado nas funções de autocorrelação (FAC) e autocorrelação parcial (FACP). Deste modo, a ordem do modelo para a parte autorregressiva, p e P , é escolhida baseada no número de defasagens estatisticamente significativas na função de autocorrelação parcial ao passo que, a escolha das ordens para a parte de médias-móveis, q e Q , são obtidas baseadas no número de defasagens significativas na função de autocorrelação. Para a identificação das ordens sazonais, P e Q , deve-se observar defasagens múltiplos de s (Morettin; Toloi, 2006).

Além disso, alguns fatores externos podem influenciar no comportamento de uma série temporal. Tais fatores podem mudar, de alguma forma, sua trajetória em algum

momento no tempo, sendo que esta mudança pode ser temporária ou permanente. Quando este tipo de evento ocorre com a série, uma análise de intervenção é necessária (Morettin; Toloi, 2006). Neste trabalho, o método de intervenção utilizado foi através da função impulso que, segundo o modelo SARIMA(p,d,q)(P,D,Q)_s, apresentado na equação (4), assume a forma

$$Z_t \begin{cases} \frac{\theta(B^s)\theta(B)}{\phi(B^s)(1-B^s)^D(1-\frac{B}{L})^d} e_t + \omega_0 I_t^{(T)} & \text{se } t = T \\ \frac{\theta(B^s)\theta(B)}{\phi(B^s)(1-B^s)^D(1-\frac{B}{L})^d} e_t & \text{se } t \neq T \end{cases} \quad (5)$$

Neste caso, o efeito de intervenção com a função impulso apresenta um comportamento atípico e temporário na série somente no instante de tempo T .

Após a estimação do modelo, espera-se que os resíduos obtidos sejam não correlacionados, ou seja, um ruído branco. Para testar a hipótese nula de que os resíduos seguem um ruído branco, utilizou-se o teste de Box-Pierce, proposto por Box e Pierce (1970). De acordo com Morettin e Toloi (2006), tomando a estatística

$$Q(K) = n(n+2) \sum_{j=1}^K \frac{\hat{r}_j^2}{(n-j)}, \quad (6)$$

tem-se que $Q(K) \xrightarrow{d} \chi_{K-p-q-P-Q}^2$, em que K é o número de defasagens utilizadas, p , q , P e Q são parâmetros do modelo SARIMA. Assim, o teste irá rejeitar H_0 se a estatística $Q(K)$ apresentar valores elevados MORETTIN; TOLOI, (2006).

Para o diagnóstico dos modelos, cujos resíduos são ruído branco, as estatísticas Critério de Informação de Akaike (AIC) e Erro Quadrático Médio de Previsão (EQMP) foram utilizadas.

O valor do AIC é dado por:

$$AIC(m) = \ln(\hat{\sigma}_a^2) + \frac{2m}{N}$$

em que $\hat{\sigma}_a^2$ é o estimador de máxima verossimilhança da variância residual do modelo considerado. Por esse critério, o melhor modelo ajustado é aquele que apresenta o menor valor para o AIC.

O modelo escolhido para o ajuste será utilizado para fazer previsões sobre observações futuras. Segundo Morettin e Toloí (2006), a previsão de Z_{t+h} para $h = 1, 2, \dots$, é definida como a esperança condicional de Z_{t+h} dados todos os valores passados. Desse modo,

$$\hat{Z}_t(h) = E[Z_{t+h} | Z_t, Z_{t-1}, \dots]$$

Essa equação será utilizada para prever o número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 2019 e dezembro de 2020.

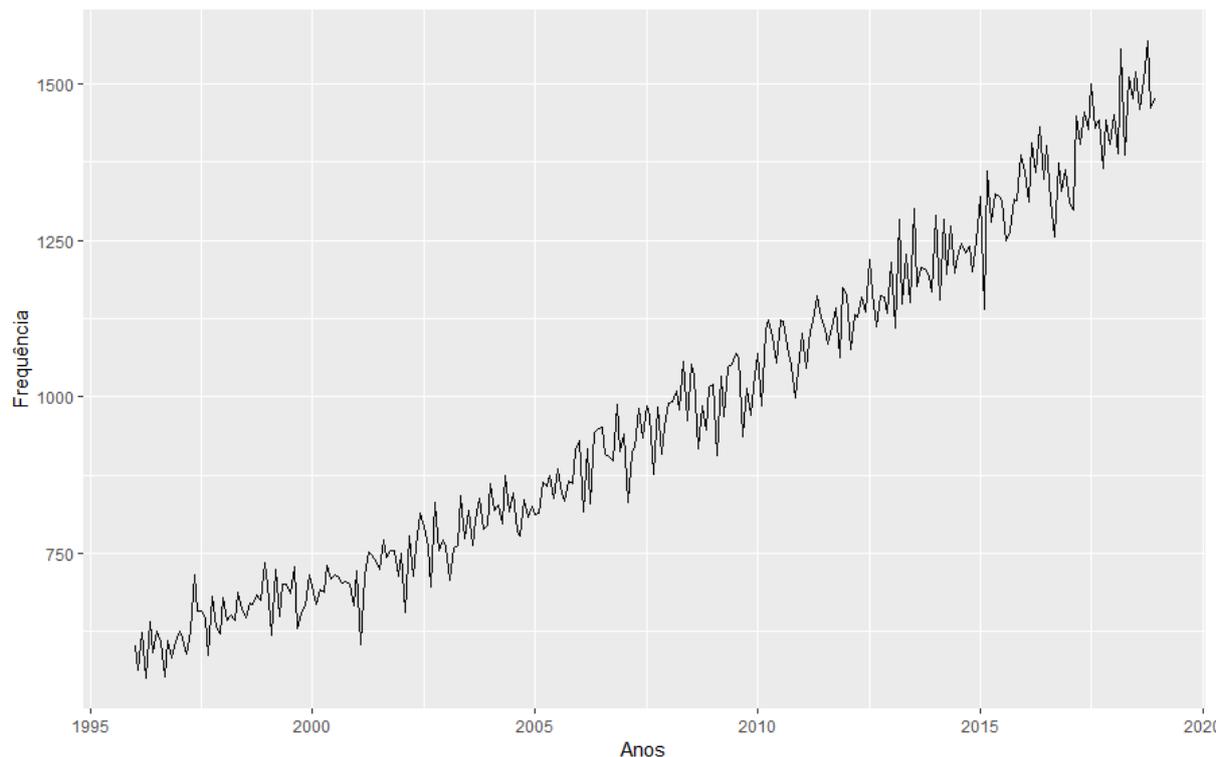
Assim, retirando as doze últimas observações da série e ajustando o modelo apropriado, foi possível calcular (EQM). Deste modo, o cálculo assumiu a seguinte forma: $EQM = \frac{1}{12} \sum_{i=1}^{12} (Z_{t+i} - \hat{Z}_{t+i})^2$. Por este critério, foi escolhido como o melhor modelo, dentre os ajustados, aquele que apresentou menor EQM.

Todo o processo de análise deste trabalho, seja construção de gráficos ou resultados de testes e ajustes, foi conduzido utilizando o *software* R (R CORE TEAM, 2020).

3. Resultados e Discussão

Na Figura 1 apresenta-se o gráfico que descreve o comportamento da série do número de óbitos por residência segundo o Ano/mês do óbito no Brasil no período de janeiro de 1996 a dezembro de 2018. Pela análise visual há indícios que exista tendência, sendo necessário a aplicação do teste de Cox-Stuart (Morettin; Toloí, 2006).

Figura 1: Série mensal de óbitos por câncer de mama no Brasil no período de janeiro de 1996 a dezembro de 2018.



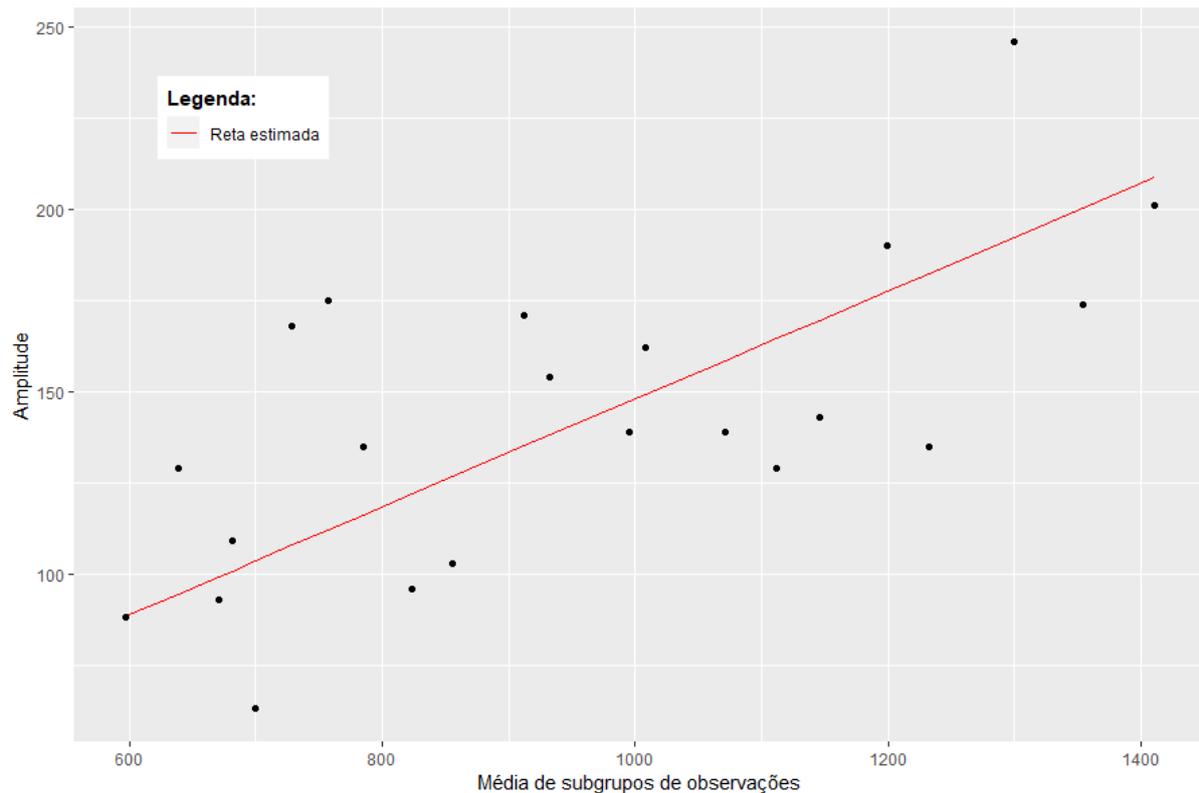
Fonte: Autores.

A fim de verificar a qualidade das previsões obtidas a partir do modelo proposto para a série, foram retiradas as doze últimas observações da série, que correspondem as observações referentes ao número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 2018 a dezembro de 2018.

Para verificar a necessidade de transformação dos dados, construiu-se o gráfico de amplitude-média Figura 2.

De acordo com Jenkins (1979), o comportamento dos pontos do gráfico indica que há necessidade de realizar transformação nos dados, visto que existe relação proporcional da amplitude com a média.

Figura 2: Gráfico da amplitude versus média de subgrupos da série temporal para verificar a necessidade de transformação nos dados.



Fonte: Autores.

Analisando o gráfico na Figura 2 foi possível observar que existe uma relação positiva entre a média de subgrupos de observações e sua respectiva amplitude, sendo que a taxa de variação, estimada por um modelo de regressão linear simples, foi de 0,148037. Neste caso, foi realizado um teste T de Student a fim de avaliar, estatisticamente se o coeficiente angular desta reta ajustada, de fato, é significativo. Assim, obteve-se um p -valor $< 0,0001$, rejeitando a hipótese nula de que o coeficiente angular da reta é nulo, confirmando a necessidade de transformação Box-Cox nos dados. Realizada a transformação, estimou-se um $\lambda = 0,367454$ da Equação (2).

Pelo teste de Cox-Stuart, observou-se a presença da componente tendência determinística com um p -valor $< 0,0001$, rejeitando, ao nível de 5% de significância, a hipótese nula de não existência de tendência. Sendo assim, foi realizada uma diferenciação de ordem 1 na série transformada a fim de remover o efeito de tendência.

Da mesma forma, foi realizado um ajuste de regressão considerando *dummies* sazonais a fim de verificar a presença de um efeito sazonal determinístico na série diferenciada.

Constatou-se que todos os meses apresentaram coeficientes significativos, ao nível de 5% de significância, confirmando a presença de sazonalidade na série.

A Tabela 1 apresenta as estimativas obtidas para a série temporal em estudo. Com base nessas estimativas, é possível observar que os meses de março, maio, outubro, julho, janeiro e dezembro foram os que apresentaram estimativas positivas, sendo março o mês mais expressivo, contribuindo para que haja um aumento no número de casos de óbitos por câncer de mama no Brasil. Já os demais meses apresentaram um padrão de decréscimo do número de casos, sendo o mês de fevereiro o mais expressivo. No entanto, não foi encontrado na literatura nenhuma explicação biológica que possa explicar este padrão sazonal observado neste conjunto de dados.

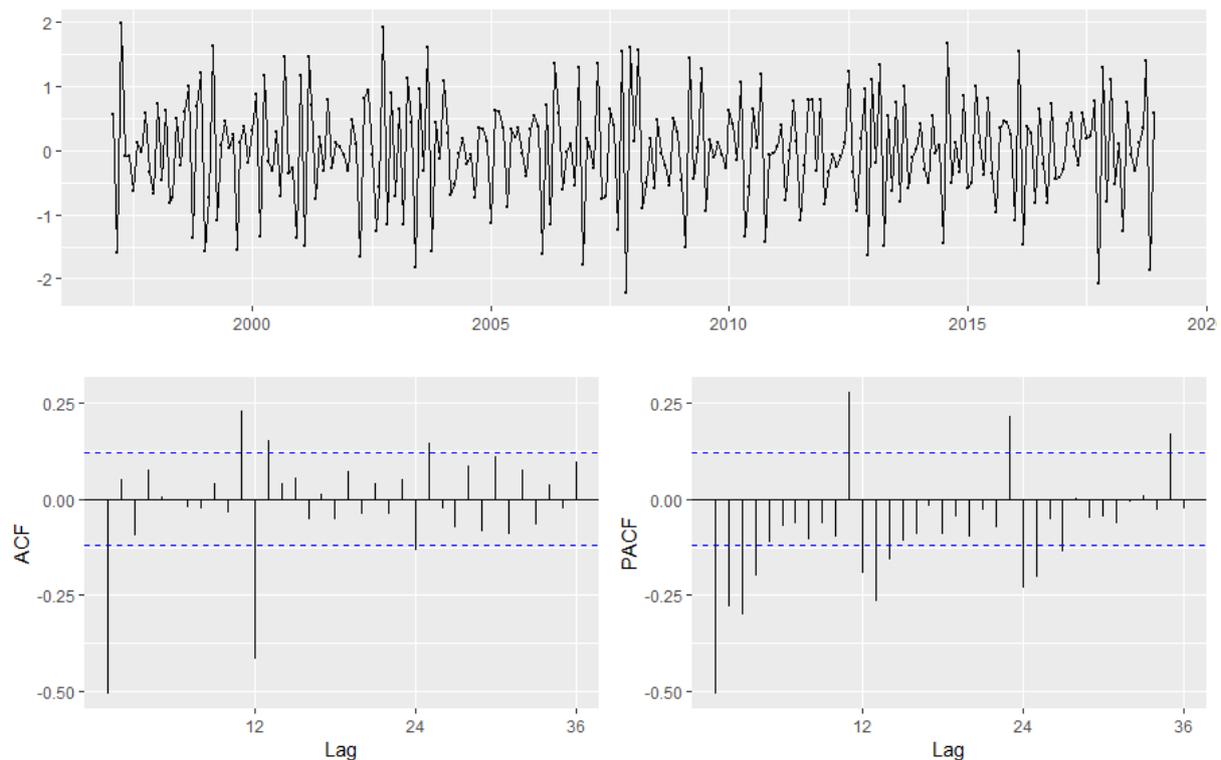
Tabela 1: Estimativas das *dummies* sazonais obtidas a partir da série temporal com transformação Box-Cox e com uma diferenciação.

Dummies sazonais	Estimativas	Erro padrão	<i>p</i> -valor
Janeiro	0.3345	0.1275	0,0092
Fevereiro	-0.9411	0.1246	<0,0001
Março	1.0791	0.1246	<0,0001
Abril	-0.4127	0.1246	0,0011
Maiο	0.7591	0.1246	<0,0001
Junho	-0.4483	0.1246	0,0004
Julho	0.3705	0.1246	0,0032
Agosto	-0.3135	0.1246	0,0125
Setembro	-0.4665	0.1246	0,0002
Outubro	0.5538	0.1246	<0,0001
Novembro	-0.2870	0.1246	0,0220
Dezembro	0.2614	0.1246	0,0369

Fonte: Autores.

Para tornar a série estacionária, foi necessário fazer uma diferença de ordem 12 na série transformada e diferenciada. Na Figura 3, apresentam-se a série de óbitos transformada e com uma diferenciação de ordem 1 e de ordem 12, bem como sua função de autocorrelação (ACF) e autocorrelação parcial (PACF).

Figura 3: Gráfico da série temporal transformada com uma diferenciação de ordem 1 e suas respectivas funções de autocorrelação e autocorrelação parcial.



Fonte: Autores.

Com base nos gráficos ACF e PACF apresentados na Figura 3, foram testados vários modelos que poderiam ser adequados para descrever a série temporal em estudo, buscando aquele com maior parcimônia e que apresentaram um comportamento de ruído branco nos resíduos.

Dentre os modelos que foram propostos, o que apresentou melhor ajuste foi o modelo $SARIMA(0,1,1)(2,1,0)_{12}$ de acordo com os critérios AIC e BIC e, portanto, este foi o modelo escolhido para descrever os dados.

Para a verificação de possíveis pontos com intervenção na série temporal, foram analisados os resíduos estimados do modelo escolhido. Assim, foi proposto um modelo $SARIMA(0,1,1)(2,1,0)_{12}$ com intervenções nos meses de fevereiro de 2001, janeiro de 2002, fevereiro de 2004, novembro de 2006 e fevereiro de 2008.

O INCA (Instituto Nacional de Câncer) ampliou as ações para a oferta de serviços no período de 1999 a 2001, resultando na realização de oito milhões de exames cito patológicos por ano. Em 2002, o fortalecimento e a qualificação da rede de atenção básica e a ampliação de centros de referência possibilitou a realização de uma segunda fase de

intensificação. O que provavelmente ocasionou o aumento do número de mortes por câncer de mama nestes períodos. O número de óbitos por neoplasias vem avançando com uma taxa média de crescimento de ao ano entre 2008 e 2016 (Tomazelli; Silva, 2017)

Na Tabela 2 são apresentadas as estimativas dos parâmetros do modelo, considerando o ajuste com e sem intervenção, obtidas pelo método da máxima verossimilhança, com seus respectivos erros padrão. Pode-se observar que todas as estimativas foram significativas.

Tabela 2: Estimativas dos parâmetros dos modelos SARIMA(0, 1, 1)(2, 1, 0)₁₂ sem e com intervenção para a série do número de óbitos por câncer de mama por residência segundo o Ano/mês do obtido no Brasil no período de Janeiro de 1996 a Dezembro de 2018.

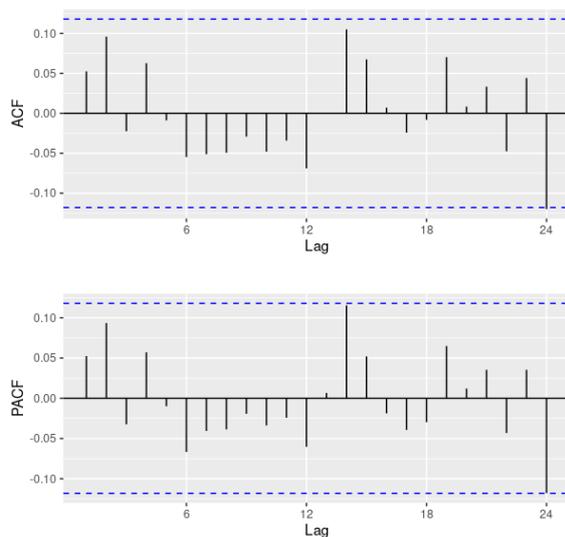
Modelo	Parâmetro	Estimativa	Erro Padrão	p-valor
SARIMA(0, 1, 1)(2, 1, 0) ₁₂ sem intervenção	Φ_1	-0,6167	0,0599	0,0000
	Φ_2	-0,3467	0,0613	0,0000
	θ_1	-0,9488	0,0285	0,0000
SARIMA(0, 1, 1)(2, 1, 0) ₁₂ com intervenção	Φ_1	-0,6342	0,0610	0,0000
	Φ_2	-0,3867	0,0616	0,0000
	θ_1	-0,9464	0,0250	0,0000
	ω_1	-1,2447	0,4118	0,0003
	ω_2	1,1073	0,4132	0,0020
	ω_3	1,4975	0,3909	0,0001
	ω_4	1,1462	0,3976	0,0054
	ω_5	-1,2410	0,4271	0,0408

Fonte: Autores.

Para o modelo SARIMA(0,1,1)(2,1,0)₁₂ sem intervenção, observou-se na Figura 4, que aproximadamente todas as autocorrelações em diferentes defasagens (“lags”) se mantiveram dentro dos limites de confiança e foram não significativos, confirmando os

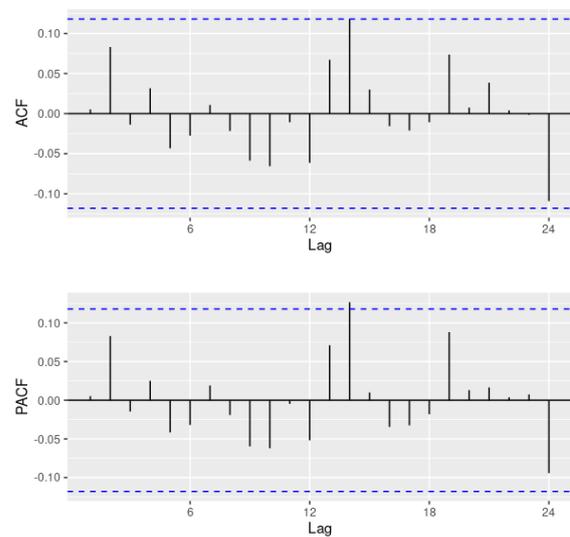
resultados apresentados pelo teste de Box e Pierce na Tabela 2 de que os resíduos deste modelo apresentaram um comportamento de ruído branco. Resultados semelhantes também foram encontrados para o modelo $SARIMA(0,1,1)(2,1,0)_{12}$ considerando intervenção, em que seu respectivo resíduo também segue um ruído branco. Tal comportamento pode ser visto conforme ilustrado nas Figuras 4 e 5.

Figura 4: Gráfico da função de autocorrelação (ACF) e autocorrelação parcial (PACF) residual do modelo $SARIMA(0,1,1)(2,1,0)_{12}$ sem intervenção.



Fonte: Autores.

Figura 5: Gráfico da função de autocorrelação (ACF) e autocorrelação parcial (PACF) residual do modelo $SARIMA(0,1,1)(2,1,0)_{12}$ com intervenção.



Fonte: Autores.

Dentre os modelos ajustados, a escolha do melhor modelo de previsão foi feita com base nos critérios de AIC, BIC e o EQMP. Na Tabela 3 são apresentados os valores obtidos através desses critérios para cada modelo e, de acordo com esses resultados, o modelo $SARIMA(0,1,1)(2,1,0)_{12}$ sem intervenção foi o modelo escolhido para representar a série do número de óbitos por câncer de mama no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2018, uma vez que apresentou o menor erro quadrático médio de previsão (EQMP).

Tabela 3: Estatística para o teste de Box-Pierce do modelo SARIMA(0, 1, 1)(2, 1, 0)₁₂ sem e com intervenção da série de observações do número de óbitos por câncer de mama no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2018.

Modelos	AIC	BIC	EQMP
SARIMA(0, 1, 1)(2, 1, 0) ₁₂ - SI	406,2694	420,5580	0,2090
SARIMA(0, 1, 1)(2, 1, 0) ₁₂ - CI	366,5433	398,6927	0,2201

Fonte: Autores.

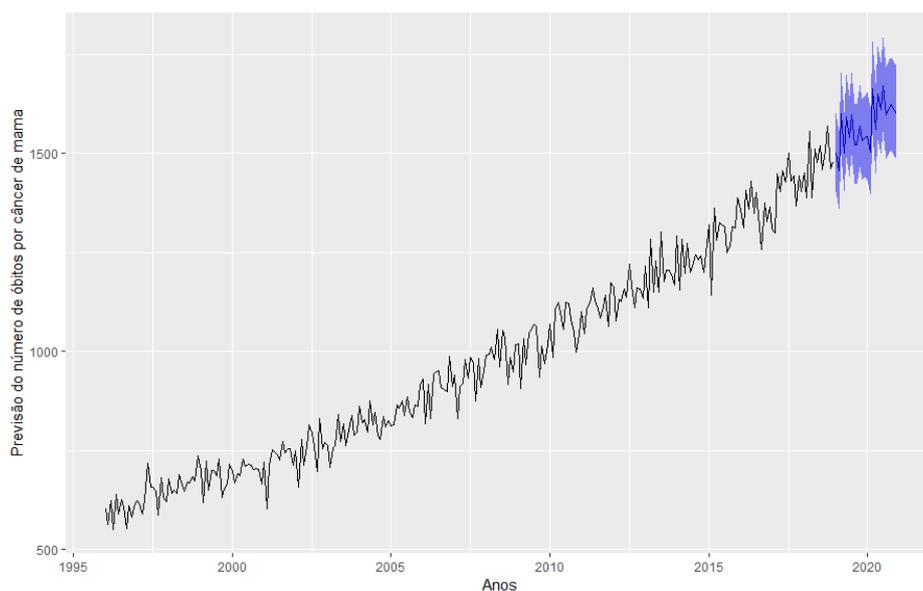
Deste modo, o modelo SARIMA(0,1,1)(2,1,0)₁₂ ajustado pode ser escrito como:

$$Z_t = \frac{(1 + 0,9488B)}{(1 - B^{12})(1 - B)(1 + 0,6167B^{12} + 0,3467B^{24})} e_t \quad (7)$$

em que e_t é um erro aleatório independente e identicamente distribuído.

A Figura 6 e a Tabela 4 apresentam as previsões do número mensal de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, entre o período de janeiro de 2019 e dezembro de 2020, como também seus respectivos intervalos com 95% de confiança, obtidas a partir do modelo SARIMA(0,1,1)(2,1,0)₁₂ sem intervenção.

Figura 6: Previsão da série transformada do número de óbitos mensais por câncer de mama no Brasil, abrangendo o período de janeiro de 2019 a dezembro de 2020.



Fonte: Autores.

Tabela 4: Previsão do número de óbitos por câncer de mama no Brasil para os anos 2019 e 2020.

Mês/Ano	Previsão	LI95%	LS95%
Jan/2019	1500.2488	1402.9722	1601.6607
Fev/2019	1455.2173	1359.6874	1554.8610
Mar/2019	1598.7902	1497.2339	1704.5722
Abr/2019	1501.7976	1404.0826	1603.6816
Mai/2019	1592.8313	1491.2572	1698.6490
Jun/2019	1541.0621	1441.4755	1644.8674
Jul/2019	1597.9126	1495.8738	1704.2206
Ago/2019	1521.6845	1422.6466	1624.9491
Set/2019	1521.3765	1422.2264	1624.7639
Out/2019	1567.9306	1466.7327	1673.4099
Nov/2019	1531.8140	1431.9811	1635.9135
Dez/2019	1539.5784	1439.2984	1644.1416
Jan/2020	1541.4293	1432.4523	1655.4772
Fev/2020	1502.5388	1395.1065	1615.0296
Mar/2020	1662.1927	1547.3900	1782.2084
Abr/2020	1559.6002	1449.1445	1675.2051
Mai/2020	1650.0221	1535.2996	1769.9904
Jun/2020	1609.8707	1496.7113	1728.2639
Jul/2020	1671.0991	1554.9872	1792.5166
Ago/2020	1598.1176	1485.0414	1716.4599
Set/2020	1615.8029	1501.7080	1735.2000
Out/2020	1622.3733	1507.7600	1742.3156
Nov/2020	1607.5221	1493.3574	1727.0244
Dez/2020	1599.9070	1485.8673	1719.2989

Fonte: Autores.

De modo geral, é possível observar, através da Figura 6 e da Tabela 4, que para os meses subsequentes a série manterá o padrão que ela vem mantendo desde o início do seu período de observação: um crescente aumento no número de óbitos por tal enfermidade.

4. Conclusão

O objetivo deste trabalho foi ajustar um modelo para série de observações do número de óbitos por câncer de mama (grupo CID-10: Neoplasias malignas na mama) no Brasil, abrangendo o período de janeiro de 1996 a dezembro de 2017. Além disso, objetivou-se prever ocorrências futuras da série para o ano de 2020.

A princípio, foi constatado que a série apresenta uma componente sazonal com periodicidade de 12 meses. Ou seja, as ocorrências de morte por câncer de mama tendem a apresentar um comportamento cíclico anualmente. Assim, um modelo sazonal foi utilizado para modelar o número de mortes. Verificou-se que o modelo $SARIMA(0,1,1)(2,1,0)_{12}$ se adequou bem aos dados, baseados nos critérios AIC e EQM. Além disso, a modelagem forneceu projeções que revelam uma queda na quantidade de mortes por câncer de mama no ano de 2017. Com isso, espera-se que as informações obtidas por meio deste trabalho possam contribuir, de alguma forma, para a melhoria na qualidade do conhecimento acerca do comportamento da desta doença no Brasil. Conhecer a forma de desenvolvimento do câncer é de suma importância, a fim de que os órgãos públicos possam alocar recursos e campanhas de conscientização e diagnóstico precoce da doença no Brasil.

Com base neste estudo, propõe-se que novos trabalhos sejam realizados através do ajuste de modelos para séries temporais considerando os dados atualizados a fim de contribuir para o desenvolvimento e efetivação de novas estratégias para o controle da doença e para uma melhor estruturação da rede de atenção à saúde. Além disso, uma outra proposta seria utilizar modelos de séries temporais considerando dados de contagem, tais como o modelo GLARMA, o que representaria melhor o conjunto de dados em análise.

Agradecimentos

Os autores Rafael Agostinho Ferreira e Patricia Mendes dos Santos agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento para este estudo e durante o período de doutorado.

Referências

Barbosa, I. R., Costa, I. do C. C., Bernal Pérez, M. M., & Souza, D. L. B. de. (2015). Mortalidade Por Câncer De Mama Nos Estados Do Nordeste Do Brasil: Tendências Atuais E

Projeções Até 2030. *Revista Ciência Plural*, 1(1), 04-14. Recuperado de <https://periodicos.ufrn.br/rcp/article/view/7318>

Batiston, A., Tamaki, E., dos Santos, M., & Cazola, L. (2009). Método De Detecção Do Câncer De Mama E Suas Implicações. *Cogitare Enfermagem*, 14(1). doi:<http://dx.doi.org/10.5380/ce.v14i1.14103>

Box, G. E. P., & Jenkins, G. M (1976). *Time Series Analysis: Forecasting and Control* (Revised). Oakland: Holden-Day

Box, G., & Pierce, D. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65(332), 1509-1526. doi:10.2307/2284333

Cox, D., & Stuart, A. (1955). Some Quick Sign Tests for Trend in Location and Dispersion. *Biometrika*, 42(1/2), 80-95. doi:10.2307/2333424

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

Padilha, P. C., & Pinheiro, R. L (2004). O papel dos alimentos funcionais na prevenção e controle do câncer de mama. *Revista brasileira de cancerologia*, 50(3), 251–260

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka R. (2018). *Metodologia da pesquisa científica. [e-book]*. Santa Maria. Ed. UAB/NTE/UFSM. Recuperado de: https://repositorio.ufsm.br/bitstream/handle/1/15824/Lic_Computacao_Metodologia-Pesquisa-Cientifica.pdf?sequence=1

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: Austria. Recuperado de <http://www.R-project.org>

Ross, S. M (1996). *Stochastic Processes* (2a ed.). New York: John Wiley & Sons

Silva, Lucia Cecília da. (2005). Vozes que contam a experiência de viver com câncer. *Psicologia Hospitalar*, 3(1), 1-17. Recuperado de http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-74092005000100002&lng=pt&tlng=pt

Tomazelli, J. G., & Silva, G. A. (2017). Rastreamento do câncer de mama no Brasil: uma avaliação da oferta e utilização da rede assistencial do Sistema Único de Saúde no período 2010-2012. *Epidemiologia e Serviços de Saúde*, 26(4), 713-724. <https://doi.org/10.5123/s1679-49742017000400004>

Porcentagem de contribuição de cada autor no manuscrito

Rafael Agostinho Ferreira – 25%

Vânia de Fátima Lemes de Miranda – 25%

Patrícia Mendes dos Santos – 20%

Henrique José de Paula Alves – 20%

Thelma Sáfadi – 10%