Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

MORTALIDADE POR ACIDENTES DE TRÂNSITO NO BRASIL: UMA ANÁLISE DE SÉRIE TEMPORAL

Arthur Oliveira Cardoso

Uberlândia-MG
2024

Arthur Oliveira Cardoso

MORTALIDADE POR ACIDENTES DE TRÂNSITO NO BRASIL: UMA ANÁLISE DE SÉRIE TEMPORAL

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Nádia Giaretta Biase

Uberlândia-MG

2024

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Nádia Giaretta Biase

_____

Pedro Franklin Cardoso Silva

_____

Maria Imaculada de Sousa Silva

Uberlândia-MG

2024

Agradecimentos

Expresso minha gratidão aos meus pais, Francisco e Edcléia, cujo constante apoio e estímulo foram fundamentais para minha trajetória acadêmica, possibilitando-me dedicar aos estudos, mesmo distante da minha cidade natal.

Aos meus amigos, que sempre estiveram ao meu lado em todos os momentos, em especial à Gabriela, que compartilhou toda a jornada da graduação comigo, tornando-a uma experiência ainda mais enriquecedora.

À minha orientadora, Nádia Giaretta Biase, pela orientação, ensinamentos, correções, incentivos e pelo comprometimento dedicado à elaboração deste trabalho.

Aos professores Pedro Franklin Cardoso Silva e Maria Imaculada de Sousa Silva, por aceitarem o convite para compor a banca examinadora e dar suas contribuições.

Resumo

Os acidentes de trânsito são uma questão grave de saúde pública em escala global, decorrentes de uma combinação de fatores comportamentais, deficiências na segurança dos veículos e condições urbanas precárias. Eles representam uma das principais causas de lesões e fatalidades, impulsionados pelo aumento contínuo do número de veículos, mudanças nos estilos de vida e comportamentos de risco observados na população em geral. Por essa razão, este estudo investiga a série temporal mensal de mortalidade por acidentes de trânsito no Brasil, com o objetivo de identificar um modelo capaz de realizar previsões para essa variável, utilizou-se técnicas de séries temporais, através da abordagem de Box e Jenkins, com modelos autorregressivos sazonais do tipo SARIMA. Os critérios adotados para tomar uma decisão analítica sobre o modelo mais apropriado incluíram o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC). Além disso, para aumentar a precisão da análise, também foram calculadas a Média Aritmética dos Desvios Absolutos em Porcentagem (MAPE) e a Raiz Quadrada da Média Aritmética dos Quadrados dos Desvios (RMSE). O modelo escolhido foi o SARIMA$(2,1,1)(3,1,1)_{12}$, que apresentou o menor valor para esses critérios.

Palavras-chave: SARIMA, Séries temporais, AIC, BIC, Acidentes de trânsito.

Abstract

Traffic accidents are a serious global public health issue, resulting from a combination of behavioral factors, deficiencies in vehicle safety, and poor urban conditions. They represent one of the main causes of injuries and fatalities, driven by the continuous increase in the number of vehicles, changes in lifestyles, and observed risk behaviors in the general population. For this reason, this study investigates the monthly time series of traffic accident mortality in Brazil, aiming to identify a model capable of making predictions for this variable. Time series techniques were employed, using the Box-Jenkins approach, with seasonal autoregressive integrated moving average (SARIMA) models. The criteria adopted to make an analytical decision about the most appropriate model included the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Additionally, to increase the accuracy of the analysis, the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) were also calculated. The chosen model was SARIMA$(2,1,1)(3,1,1)_{12}$, which presented the lowest value for these criteria.

Keywords: SARIMA, Time series, AIC, BIC, Traffic accidents.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Os acidentes de trânsito representam uma séria questão de saúde pública em todo o mundo, sendo resultado de fatores comportamentais, falta de segurança nos veículos e condições urbanas precárias. Eles figuram como as principais razões por trás de lesões e fatalidades, devido ao aumento constante no número de veículos, mudanças nos padrões de vida e comportamentos de risco observados na população em geral[13].

Globalmente, entre 20 e 50 milhões de pessoas são vítimas de lesões e incapacitações anualmente em decorrência de acidentes de trânsito, resultando em 1,3 milhão de mortes. Os dados mostram que as lesões causadas pelo trânsito continuam a ser um importante problema de saúde pública, especialmente para países de baixa e média renda[22]. O Brasil ocupa a terceira posição nesse trágico ranking, ficando atrás apenas da China e Índia[4]. Pesquisas epidemiológicas têm evidenciado que os acidentes de trânsito exibem padrões distintos de distribuição conforme o sexo, idade, estratos sociais e regiões de risco. Isso revela situações de vulnerabilidade tanto para indivíduos quanto para determinadas localidades[13].

Além da tragédia irreparável de perder familiares e entes queridos de forma súbita, os acidentes de trânsito no Brasil também resultam em um enorme ônus econômico suportado pelo governo e, consequentemente, por toda a sociedade. De acordo com um estudo realizado pelo aplicativo Gringo em parceria com o Centro de Liderança Pública (CLP), estimou-se que, anualmente, perdemos cerca de R$ 20 bilhões devido às mortes nas vias. Esse valor representa 0,2% do Produto Interno Bruto (PIB) do país em 2022. No período de 2012 a 2020, ocorreram 335.424 óbitos. O custo associado a essas vidas perdidas em acidentes de trânsito foi calculado em R$ 180 bilhões, com cada morte representando, em média, R$ 536 mil[20].

Em 2010, a Organização das Nações Unidas (ONU) lançou a campanha "1ª Década de Ação pela Segurança no Trânsito", com o objetivo de sensibilizar os países a adotarem medidas para reduzir em 50% a mortalidade no trânsito até 2020. Contudo, em agosto de 2023, o Instituto de Pesquisa Econômica Aplicada (Ipea) divulgou o relatório "Balanço da 1ª década de ação pela segurança no trânsito no Brasil e perspectivas para a 2ª década". O estudo revelou que entre 2010 e 2019, o Brasil registrou um aumento de 13,5% no número absoluto de mortes no trânsito em comparação com a década anterior. Além disso, a taxa de mortalidade por 100 mil habitantes cresceu 2,3% nesse período, indicando resultados decepcionantes em relação à meta global estabelecida pela ONU[10].

Para estudar as taxas de mortalidade em um determinado momento do passado e tentar prever o comportamento futuro, os métodos de séries temporais são ferramentas eficazes. Essas

técnicas têm o potencial de fornecer resultados promissores, podendo contribuir parcialmente para resolver esses problemas. Elas oferecem informações de qualidade que podem auxiliar os gestores municipais, estaduais e federais a desenvolver estratégias de intervenção eficazes. Além disso, ajudam a direcionar investimentos e a elaborar políticas públicas voltadas para a redução ou eliminação de problemas, danos e mortes prematuras.

Nesse sentido, o objetivo desse estudo foi modelar a série temporal mensal de mortalidade por acidentes de trânsito no Brasil, por meio dos modelos de Box e Jenkins, e encontrar o modelo apropriado que apresenta o melhor resultado de previsão. A partir dele, caso seja possível, pretende-se indicar os períodos do ano que esses eventos são mais recorrentes, afim de fornecer subsídios para a elaboração de políticas públicas destinadas a reduzir tais eventos, os quais acarretam sérios prejuízos sociais, especialmente nos aspectos socioeconômicos de uma comunidade.

2. Fundamentação Teórica

2.1 Séries temporais

A suposição básica que norteia a análise de séries temporais é que há um sistema causal mais ou menos constante, relacionado com o tempo, que exerceu influência sobre os dados no passado e pode continuar a fazê-lo no futuro. Este sistema causal costuma atuar criando padrões não aleatórios que podem ser detectados em um gráfico da série temporal, ou mediante algum outro processo estatístico.

O objetivo da análise de séries temporais é identificar padrões não aleatórios na série temporal de uma variável de interesse, e a observação deste comportamento passado pode permitir fazer previsões sobre o futuro, orientando a tomada de decisões[18].

Para conduzir a análise de uma série temporal, é fundamental que os dados estejam organizados em intervalos regulares de tempo e ordenados cronologicamente. Além disso, uma série temporal pode ser caracterizada pela presença de componentes de tendência, sazonais, cíclicas e de ruído, as quais serão exploradas mais detalhadamente ao longo deste estudo.

2.1.1 Estacionariedade

Uma das suposições mais comuns em relação a uma série temporal é sua estacionariedade, indicando que ela se desenvolve aleatoriamente em torno de uma média constante ao longo do tempo, refletindo um estado de equilíbrio estável. No entanto, na prática, a maioria das séries temporais apresenta algum grau de não estacionariedade. Embora uma série possa ser estacionária por um período prolongado, é possível que ela seja estacionária apenas em intervalos muito curtos, sofrendo mudanças em seu nível e/ou inclinação ao longo do tempo.

Dado que a maioria dos métodos de análise estatística de séries temporais presume que estas sejam estacionárias, é essencial realizar transformações nos dados originais caso estes não formem uma série estacionária. Uma das transformações mais comuns envolve a aplicação de diferenças sucessivas à série original, até que uma série estacionária seja obtida. Segundo Morettin e Toloi (2006)[11], a primeira diferença da série temporal, representada por $Z(t)$, é definida por

$$\triangle Z(t) = Z(t) - Z(t-1), \tag{2.1}$$

a segunda diferença é

$$\triangle^2 Z(t) = \triangle[\triangle Z(t)] = \triangle[Z(t) - Z(t-1)], \tag{2.2}$$

ou seja,

$$\triangle^2 Z(t) = Z(t) - 2Z(t-1) + Z(t-2). \tag{2.3}$$

De modo geral, a n-ésima diferença de $Z(t)$ é

$$\triangle^n Z(t) = \triangle[\triangle^{n-1} Z(t)]. \tag{2.4}$$

Em circunstâncias típicas, geralmente é adequado aplicar uma ou duas diferenças para converter a série em uma forma estacionária.

2.1.2 Tendência e Sazonalidade

Consideraremos as observações $Z_t, t = 1, ..., n$ de uma série temporal. Um modelo de decomposição consiste em escrever Z_t como uma soma de três componentes não-observáveis,

$$Z_t = T_t + S_t + a_t, \tag{2.5}$$

em que T_t e S_t representam a tendência e a sazonalidade, respectivamente, enquanto a_t é uma componente aleatória, de média zero e variância constante ${\sigma_a}^2$.

Tendência

Segundo Morettin e Toloi (2006)[11], a tendência é caracterizada pelo aumento ou pela diminuição gradual das observações ao longo do período. Há vários métodos para estimar T_t. Os mais utilizados consistem em:

- ajustar uma função do tempo, como um polinômio, uma exponencial ou outra função suave de t;
- suavizar (ou filtrar) os valores da série ao redor de um ponto, para estimar a tendência naquele ponto;
- suavizar os valores da série através de sucessivos ajustes de retas de mínimos quadrados ponderados.

Para atestar a tendência de uma série temporal, existem diversos testes de hipóteses que podem ser utilizados após a estimação de T_t. No presente estudo, o teste utilizado foi o de Mann-Kendall.

O teste de Mann-Kendall é um teste não-paramétrico para identificar tendências em dados de séries temporais. Mann (1945) sugeriu uma comparação entre aleatoriedade e tendência, enquanto Kendall (1975) havia proposto um teste para identificar correlação, cuja aplicação

é semelhante à condição apresentada por Mann. A estatística do teste proposta por Kendall (1975) é dada por[19]:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sng(Z_j - Z_i) \tag{2.6}$$

sendo n o número de pontos de dados, Z_i e Z_j os respectivos valores dos dados em séries de tempo $i = 1, ..., n-1$ e $j = i+1, i+2, ..., n$, e $sng(Z_j - Z_i)$ dada por:

$$sng(x_j - x_i) = \begin{cases} +1, & \text{se } x_j - x_i > 0 \\ 0, & \text{se } x_j - x_i = 0 \\ -1, & \text{se } x_j - x_i < 0 \end{cases} \tag{2.7}$$

Kendall (1975) mostrou que S é normalmente distribuída com média $E(S)$ e variância $Var(S)$. A variância é calculada como:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i-1}^{m}[t_i(t_i-1)(2t_i+5)]}{18} \tag{2.8}$$

em que n é o número de pontos dos dados, m é o número de grupos empatados (conjunto de dados da amostra com valores idênticos) e t_i indica a quantidade de valores repetidos no i-ésimo grupo.

Se o tamanho da amostra n é maior que 10, a estatística de teste normal padrão Z_s é calculada da seguinte forma:

$$Z_s = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, & \text{se } S > 0 \\ 0, & \text{se } S = 0 \\ \frac{S+1}{\sqrt{Var(S)}}, & \text{se } S < 0 \end{cases} \tag{2.9}$$

Valores positivos de Z_s indicam tendências crescentes, enquanto valores negativos de Z_s denotam tendências decrescentes. A hipótese nula H_0 desse teste é de que não existe tendência na série de dados, ou seja, caso H_0 seja rejeitada, conclui-se que a série original possui tendência. As tendências de teste são realizadas ao nível de significância específico, neste trabalho será utilizado nível de significância de 5%.

Sazonalidade

A sazonalidade em uma série se refere às flutuações de ascensão e queda que ocorrem regularmente durante um determinado período do ano, mês, semana ou dia. Essa sazonalidade pode ser classificada como determinística, o que significa que pode ser prevista com precisão com base nos meses anteriores, ou estocástica, o que implica que a componente sazonal varia ao longo do tempo. A distinção fundamental entre as componentes sazonal e cíclica reside no fato de que a primeira apresenta padrões previsíveis, ocorrendo em intervalos regulares de tempo, geralmente menores do que um ano, ao passo que os movimentos cíclicos tendem a ser mais

irregulares[11].

Priestley (1989)[17] introduz o teste de Fisher para avaliar a presença de sazonalidade determinística, utilizando a análise de uma medida conhecida como periodograma, a qual é calculada com base nas funções seno e cosseno.

O periodograma descreve os valores observados em uma realização de uma série temporal através da sobreposição de ondas sinusoidais em diferentes frequências. Uma aplicação prática fundamental dessa decomposição é sua utilidade na identificação de componentes cíclicas ou periódicas.

Conforme Priestley (1989)[17], a função periódica é dada por:

$$I_p(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t cos\frac{2\pi i}{n}t\right)\left(\sum_{t=1}^{n} a_t sen\frac{2\pi i}{n}t\right)^2\right] \tag{2.10}$$

em que $0 < f_i < \frac{1}{2}$, $t = 1, 2, ..., n$ e $I_p(f_i)$ é a intensidade da frequência f_i. A periodicidade de período $\frac{1}{f_i}$ pode ser observada pela existência de picos na frequência $f_i = \frac{1}{n}$.

Nesse estudo, será utilizado o teste de Fisher para avaliar a presença de sazonalidade na série, a hipótese nula H_0 é de que não existe sazonalidade.

A estatística do teste é dada por

$$g = \frac{max I_p}{\sum_{p=1}^{n/2} I_p}, \tag{2.11}$$

em que I_p é o valor do peridiograma especificado na equação 2.10 no período p e n é o número de observações da série. A estatística do teste de Fisher, Z_α, é

$$Z_\alpha = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{\frac{n}{2}-1}} \tag{2.12}$$

sendo α o nível de significância do teste. Se $g > z$, rejeita-se H_0, ou seja, a série apresenta periodicidade p.

2.1.3 Modelo Autorregressivo Integrado de Médias Móveis (ARIMA)

Uma abordagem amplamente empregada na análise de modelos paramétricos é denominada abordagem de Box e Jenkins (1976). Essa metodologia envolve a aplicação de modelos autorregressivos integrados de médias móveis, ARIMA(p, d, q), utilizado para ajustar a um conjunto de dados de uma série temporal não estacionária[11].

De acordo com Morettin e Toloi (2006)[11], a construção do modelo segue uma estratégia baseada em um ciclo iterativo, composto pelos estágios de especificação, identificação, estimação e verificação. Na etapa de especificação, uma classe ampla de modelos é considerada para análise. Na segunda fase, o modelo é identificado através da análise da função de autocorrelação (ACF) e da função de autocorrelação parcial (PACF). Na fase de estimação do modelo, os parâmetros identificados anteriormente são estimados. A última etapa consiste em testar o

modelo, realizando um diagnóstico do modelo ajustado. Para isso, examinam-se os resíduos da regressão estimada e avalia-se a capacidade de previsão do modelo. Se o modelo for considerado adequado, pode ser utilizado para previsão; caso contrário, é necessário reiniciar todo o processo.

Conforme descrito por Fava (2000)[5], os modelos ARIMA são compostos pela combinação de três componentes denominados "filtros": o componente autorregressivo (AR), o filtro de integração (I) e o componente de médias móveis (MA). Uma série temporal pode ser modelada utilizando os três filtros em conjunto ou apenas um subconjunto deles, o que resulta em diferentes modelos, conforme abordado a seguir.

Um modelo autorregressivo é formulado de modo que os valores da série temporal no tempo t sejam influenciados pelos valores anteriores. De forma mais precisa, um modelo autorregressivo de ordem p, representado pela notação AR(p), é expresso por:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t, \tag{2.13}$$

em que:

$$\phi(B) = 1 - \phi_1(B) - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2.14}$$

é o operador autorregressivo estacionário de ordem p, e a série a_t representa o resíduo, independente e identicamente distribuído, denominado de ruído branco. Isso acontece quando as variáveis aleatórias a_t não estão correlacionadas, ao mesmo tempo em que possuem média igual a zero e uma variância constante.

Considerando um processo linear Z_t, obtemos um processo de médias móveis de ordem q, denotado por MA(q), expressado por:

$$Z_t = \mu + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{2.15}$$

e sendo $\bar{Z}_t = Z_t - \mu$, teremos:

$$\bar{Z}_t = (1 - \theta_1 B - \cdots - \theta_q B^q) a_t = \theta(B) a_t, \tag{2.16}$$

em que:

$$\theta(B) = 1 - \theta_1 B - \theta_1 B^2 - \cdots - \theta_q B^q \tag{2.17}$$

é o operador de médias móveis de ordem q.

Para várias séries observadas na prática, quando desejamos um modelo com um número limitado de parâmetros, a incorporação de termos autorregressivos e de médias móveis é uma solução apropriada.

Surgem, então, os modelos ARMA(p, q), da forma:

$$\bar{Z}_t = \phi_1 \bar{Z}_{t-1} + \cdots + \phi_p \bar{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{2.18}$$

Um processo ARMA(p, q), tem uma função de autocorrelação (ACF) infinita em extensão,

a qual decai de acordo com exponenciais e/ou senóides amortecidas após o "lag" $q-p$. E possui uma função de autocorrelação parcial (PACF) denominada por exponenciais e/ou senóides amortecidas.

Os modelos mencionados anteriormente são adequados para descrever séries temporais estacionárias, ou seja, séries que variam em torno de uma média constante ao longo do tempo. No caso de séries temporais não estacionárias, é necessário aplicar diferenças sucessivas nos dados originais até que a série se torne estacionária. Esses modelos são conhecidos como modelos ARIMA.

Se W_t for uma diferença de Z_t, então Z_t é uma integral de W_t, daí dizermos que Z_t segue um modelo autorregressivo integrado de médias móveis, ou modelo ARIMA:

$$\phi(B)\Delta^d Z_t = \theta(B)a_t, \tag{2.19}$$

de ordem (p, d, q) e escrevemos ARIMA(p, d, q), se p e q são as ordens de $\phi(B)$ e $\theta(B)$, respectivamente, e d são as diferenças sucessivas aplicadas na série original para que ela se torne estacionária.

Muitas séries temporais exibem padrões recorrentes que se manifestam regularmente em intervalos de tempo específicos, chamados de sazonalidade. Para lidar com séries temporais que exibem autocorrelação sazonal, Box e Jenkins (1976) ampliaram o modelo ARIMA, introduzindo o modelo ARIMA sazonal multiplicativo, conhecido como SARIMA$(p, d, q)(P, D, Q)_S$, e representado por[21]:

$$\phi_p(B)\Phi_p(B^s)\Delta^d\Delta_S^D Z_t = \theta_q(B)\Theta_Q(B^S)a_t \tag{2.20}$$

em que:

- $\Phi_p(B^S)$ é o operador sazonal AR(p) definido por $\Phi_p(B^S) = 1 - \Phi_1(B^S) - \Phi_2(B^{2S}) - \cdots - \Phi_p(B^{PS})$;
- $\Theta_Q(B^S)$ é o operador sazonal MA(q) definido por $\Theta_Q(B^S) = 1 - \Theta_1(B^S) - \Theta_2(B^{2S}) - \cdots - \Theta_Q(B^{QS})$;
- $\Phi_1, ..., \Phi_p$ são os parâmetros do modelo sazonal AR(p);
- $\Theta_1, ..., \Theta_Q$ são os parâmetros do modelo sazonal MA(q).
- Δ_S^d é o operador diferença simples, d indicando a ordem de diferenciação simples;
- $\Delta_S^D = (1 - B^S)^D$ é o operador diferença sazonal S, D indicando a ordem de diferenciação sazonal;

2.1.4 MEDIDAS DE COMPARAÇÃO DE MODELOS

Neste estudo, inúmeros modelos serão ajustados até que seja escolhido o que mais se adequa aos dados. Para realizar a comparação desses modelos, diversas métricas foram empregadas

para comparar os modelos até a identificação daquele mais apropriado. No decorrer deste documento, tais métricas serão detalhadas e explicadas.

Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike[2] é dado por:

$$AIC = -2logL(q) + 2p, \tag{2.21}$$

em que $L(q)$ é o máximo da função de verossimilhança, p é o número de parâmetros no modelo e log é o operador logaritmico natural.

O primeiro componente é uma recompensa por uma melhor adaptação aos dados, enquanto o segundo componente é uma penalidade. O modelo selecionado como o melhor será aquele que exibir o menor valor de AIC.

Critério de Informação Bayesiano (BIC)

O Critério de Informação Bayesiano é obtido pela seguinte expressão:

$$BIC = -2logL(q) + plog(n), \tag{2.22}$$

em que $L(q)$ é o máximo da função de verossimilhança, p é o número de parâmetros no modelo, n é o número de observações utilizadas na estimação do modelo em estudo e log é o operador logaritimico natural[12].

O critério de informação bayesiano (BIC) é uma estatística que busca maximizar a probabilidade de identificar o modelo mais adequado dentre os avaliados. Quanto menor for o valor do critério de informação, melhor será o modelo.

Média Aritmética dos Desvios Absolutos em Porcentagem (MAPE)

A Média Aritmética dos Desvios Absolutos em Porcentagem avalia a magnitude do erro com relação à série histórica, calculado da seguinte forma[3]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - p_i}{y_i}\right| \times 100 \tag{2.23}$$

onde n é o número total de amostras, y_i é o valor real para a amostra i e p_i é o valor previsto para a amostra i.

É a mais amplamente adotada entre as abordagens para avaliar a precisão das previsões (Kahn, 1998). No entanto, a aplicação desta fórmula torna-se inviável quando a série temporal inclui valores iguais a zero. Por ser um erro relativo, ele não está vinculado à escala dos dados, possibilitando a comparação da precisão das previsões entre séries temporais de proporções distintas[3].

Raiz Quadrada da Média Aritmética dos Quadrados dos Desvios (RMSE)

A Média Aritmética dos Quadrados dos Desvios (EQM), destaca os grandes erros, comparados aos erros de menor magnitudade, e é obtida pelo seguinte cálculo[3]:

$$EQM = \frac{1}{n}\sum_{t=1}^{n} a_t^2 \tag{2.24}$$

O RMSE é a raiz quadrada dos EQM's, portanto, é obtido pela raiz quadrada da equação 2.24. Esta medida também tende a exagerar erros grandes, que podem ajudar a eliminar modelos com grandes erros.

3. METODOLOGIA

O conjunto de dados a ser utilizado na presente pesquisa é o Sistema de Informação sobre Mortalidade - SIM, que está disponibilizado no site do Governo Federal[7]. Através desse conjunto de dados, será gerado a série temporal mensal de mortalidade por acidentes de trânsito no Brasil, e para a análise será considerado o período de janeiro de 2000 a dezembro de 2022, totalizando 276 meses, ou 23 anos. Para que seja possível verificar a adequação do modelo, as últimas 12 observações serão retiradas, de modo que sejam fornecidas previsões para os meses de janeiro a dezembro de 2022, e ao fim do processo será comparado os valores realizados com os previstos.

Primeiramente, o conjuto de dados completo, que inclui outras variáveis, foi avaliado para verificar uma possível influência sobre a mortalidade. Todas as informações das variáveis foram importadas para o software R, para que fosse realizada uma análise descritiva. Foram feitos gráficos, utilizando o pacote *ggplot2*[14], para analisar a mortalidade com relação a idade, o período, o sexo e a região.

Após feita a análise descritiva dos dados, foi importada a série temporal para o software R, para ser efetuado o ajuste do modelo. O R possui recursos computacionais específicos para tratar séries temporais, especialmente o pacote *forecast*[9], que é o pacote que realiza a estimação do modelo. Para realizar o ajuste do modelo, antes é necessário verificar se a série temporal satisfaz alguns pressupostos, que serão detalhados no decorrer do texto.

Para verificar a existência de tendência nos dados, foi utilizado o teste de Mann-Kendall, o qual está dispónivel no pacote *trend*[15] com o comando *mk.test()*[8]. O método baseia-se em rejeitar ou não a hipótese nula H_0, de que não exista tendência na série de dados. Caso H_0 seja rejeitada, conclui-se a série original tem tendência. Nesse caso, é necessário tomar diferenças na série para torná-la estacionária. Esse passo deve ser repetido até que não haja mais tendência nos dados, com incremento de uma diferenciação a cada repetição, mas espera-se que esse grau de diferenciação não seja maior que 2.

Na sequência, foi feita a verificação da existência de sazonalidade na série temporal. O teste de Fisher, que está presente no pacote *GeneCycle*[1] com o comando *fisher.g.test()*[6] dentro do software R, foi o utilizado para verificar a presença da componente sazonal. Sua hipótese nula H_0 afirma que não existe sazonalidade, portanto, caso esta seja rejeitada, deve-se concluir que a componente sazonal é significativa, sendo assim necessário ajustar um modelo do tipo SARIMA para os dados.

Agora, a próxima etapa é realizar o ajuste do modelo. Utilizando o comando

auto.arima()[16] presente no pacote *forecast*, foram testados diversos modelos com diferentes combinações de parâmetros, a fim de encontrar o modelo que mais se adeque aos dados. As medidas mais recomendadas para escolha do modelo são o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC). Foram calculados os valores do AIC e do BIC para cada um dos modelos, e será considerado como melhor modelo o que possuir os menores valores para essas medidas, e prosseguirá para a análise dos seus resíduos.

Prosseguindo para a análise dos resíduos, para assegurar a sua normalidade, serão aplicados os testes de autocorrelação residual, e o teste de Box-Pierce, que é utilizado para verificar se os resíduos são independentes e compõem um ruído branco.

E por fim, será realizada a predição dos dados para o ano de 2022 utilizando o modelo escolhido. O pacote *forecast* inclui um comando de previsão que, ao ser executado, gera os valores previstos. Esse resultado pode ser plotado e comparado com os valores reais.

4. RESULTADOS

Inicialmente, foi realizada uma estatística descritiva dos dados, utilizando o software R, acerca das variáveis presentes no banco de dados, tais como idade, sexo, região, dentre outras, relacionando-as com a mortalidade a fim de identificar alguma relação entre elas. Para visualizar primeiro apenas a mortalidade ao longo dos anos, foi plotado a Figura 4.1.

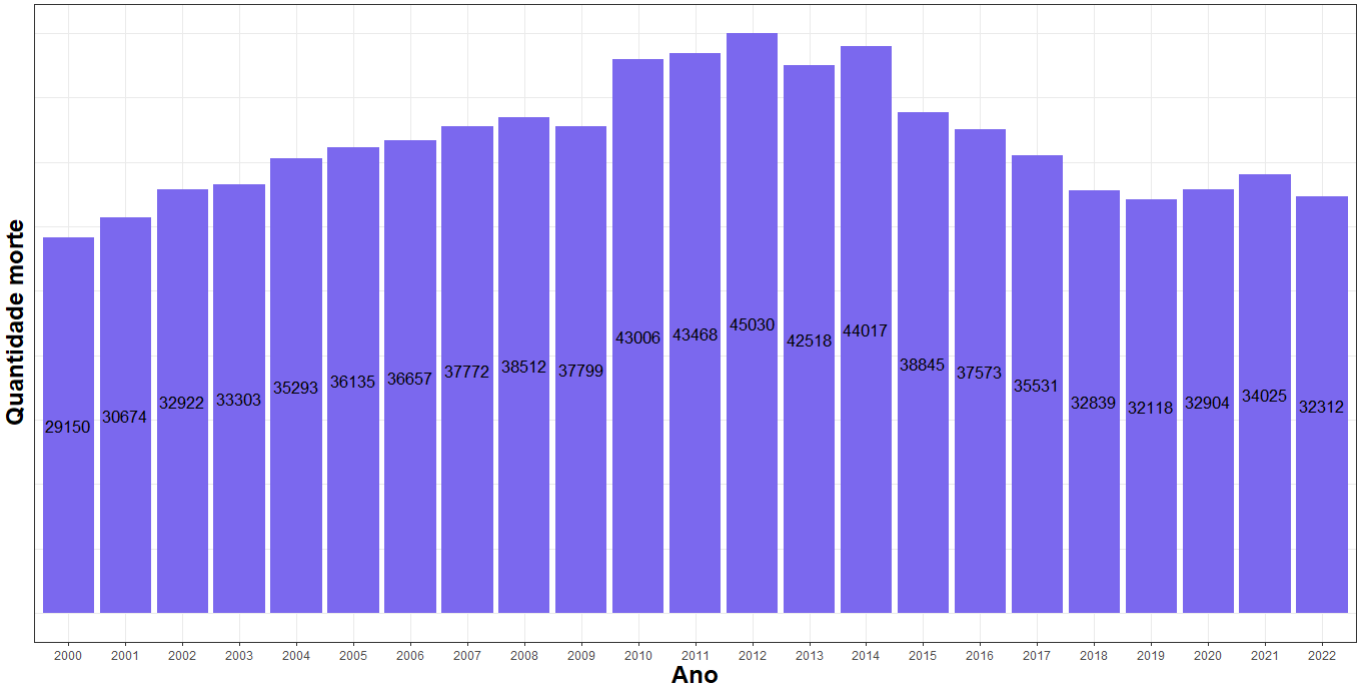

Figura 4.1: Mortalidade por acidente de trânsito no Brasil por ano.

Por meio da Figura 4.1 é possível observar um volume maior de mortalidade entre os anos de 2010 e 2014, havendo uma queda após esse período. Vale ressaltar que com a implementação da Lei Seca em 2008 é observado uma queda imeditada no ano de 2009, mas logo nos anos seguintes esse número volta a crescer, o que pode estar relacionado com a difícil fiscalização da polícia nas rodovias brasileiras.

Na Figura 4.2 é apresentada a mortalidade por acidentes em função do sexo dos indivíduos. Pode-se notar que em todos os anos a quantidade de morte de pessoas do sexo masculino é consideravelmente superior a de pessoas do sexo feminino.

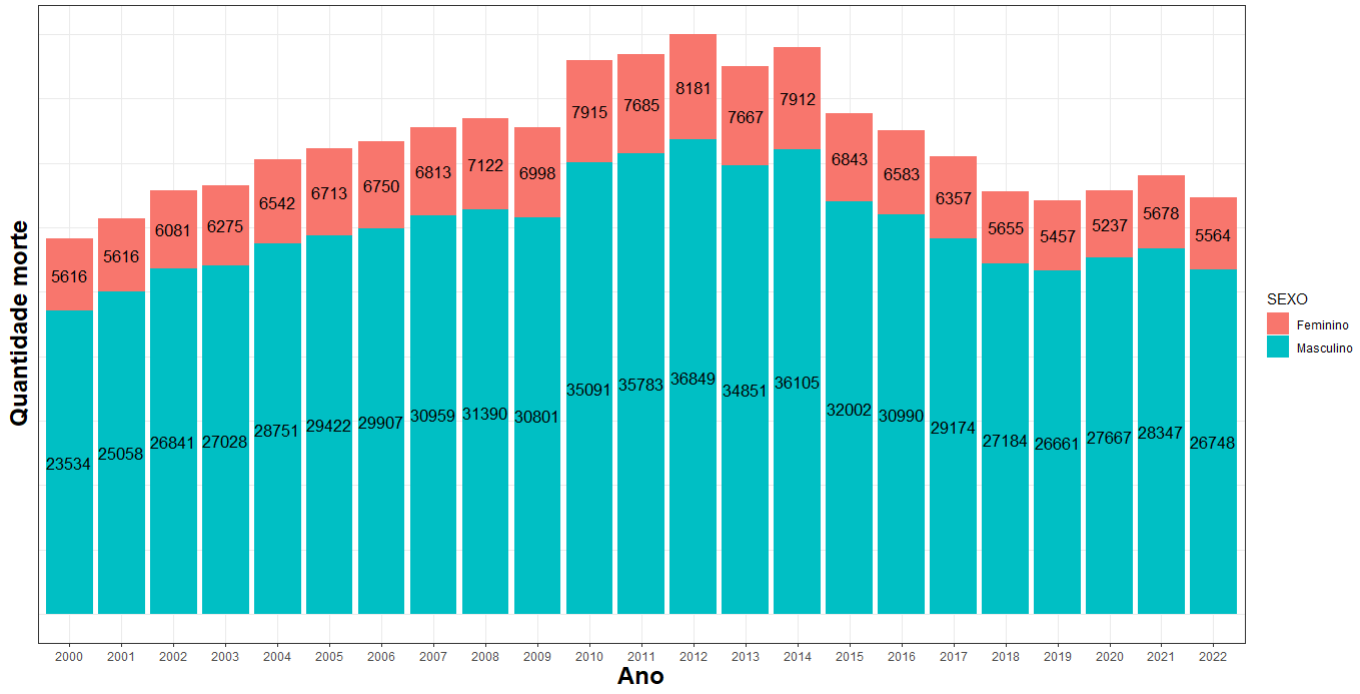

Figura 4.2: Mortalidade por acidente de trânsito no Brasil por sexo.

Quando foi plotado o gráfico relacionando a mortalidade com a região, conforme pode ser visualizado na Figura 4.3, notou-se que as regiões Nordeste e Sudeste possuem o maior volume de mortalidade, o que já era esperado pois além de serem as mais populosas do país, essas regiões recebem o maior número de turistas anualmente, destacando-se cidades como São Paulo, Rio de Janeiro, Belo Horizonte, Salvador, Fortaleza e Maceió.

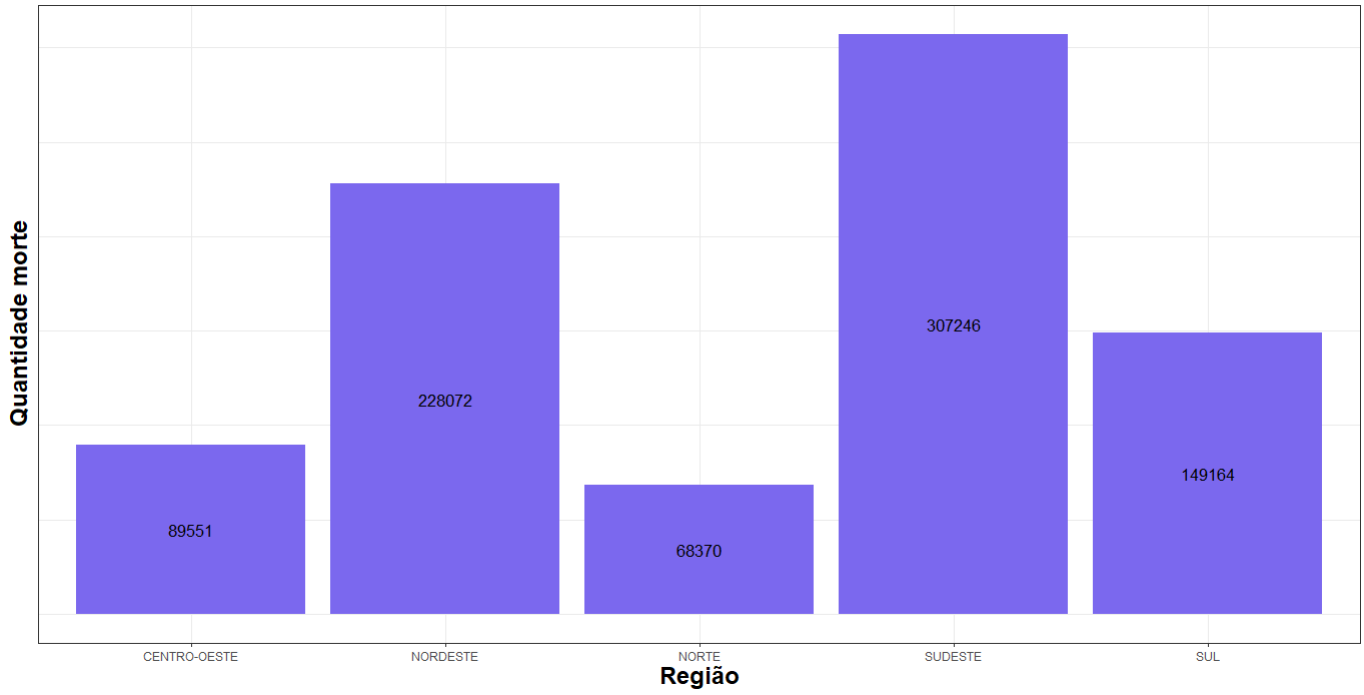

Figura 4.3: Mortalidade por acidente de trânsito no Brasil por região.

Na Figura 4.4 é apresentado o número de mortes por acidentes de trânsito em função do mês do ano. Nesse caso, foi possível visualizar que o mês de dezembro apresentou uma quantidade maior de mortalidade quando comparado com os outros meses, mesmo não sendo uma diferença muito expressiva. Esse aumento pode estar associado com as festividades de final de ano, em que um número maior de pessoas transitam em rodovias para passar esse período com parentes ou regiões litorâneas, e, consequentemente, o uso de bebidas alcóolicas também é maior, o que pode contribuir para esse acréscimo. Outro fator que pode contribuir para esse aumento são as condições climáticas nesse período do ano, onde o volume de chuvas é maior, o que afeta diretamente as condições das rodovias.

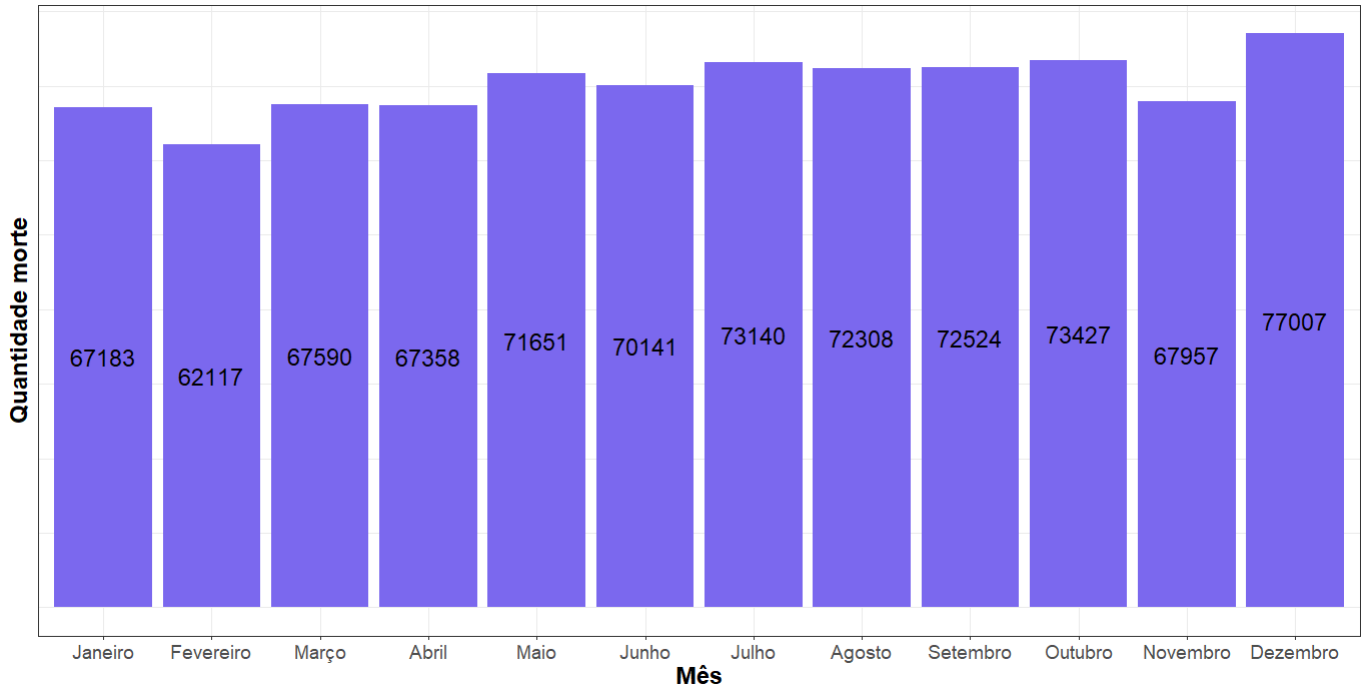

Figura 4.4: Mortalidade por acidente de trânsito no Brasil por mês.

Na Figura 4.5 é apresentado o gráfico de barras da mortalidade por acidente de trânsito em função da idade e sexo dos indivíduos. Nota-se que o maior volume de mortalidade é encontrado no grupo de pessoas de 16 a 45 anos, sendo desses a maioria pessoas do sexo masculino, destacando-se a imprudência e o comportamento do condutor associados ao consumo de bebidas alcoólicas dentre as principais causas dessa mortandade.

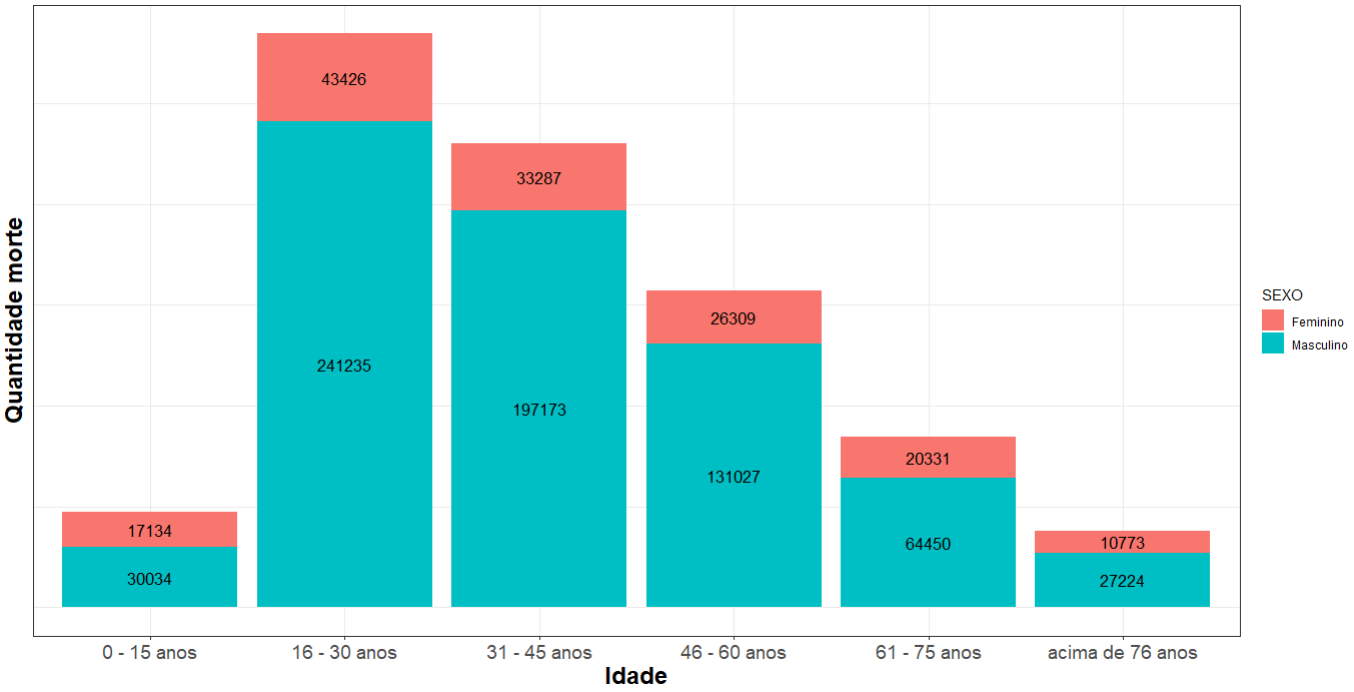

Figura 4.5: Mortalidade por acidente de trânsito no Brasil por idade.

Após a análise descritiva dos dados, foi iniciado o ajuste do modelo de série temporal. Na Figura 4.6 é ilustrada a série original da mortalidade por acidente de trânsito no Brasil, compreendida entre o período de janeiro de 2000 a dezembro de 2022. A inspeção visual dessa série nos permite avaliar o comportamento preliminar das componentes presentes nessa análise. É notório uma tendência crescente entre os anos de 2000 e 2014, como também entre 2021 e 2022. Entre os anos de 2015 e 2020, observou-se uma tendência descrescente. Além disso, a presença da sazonaliddade, também ficou bastante evidente durante todo o período em estudo. Isso nos leva a concluir que trata-se de uma série não estacionária.

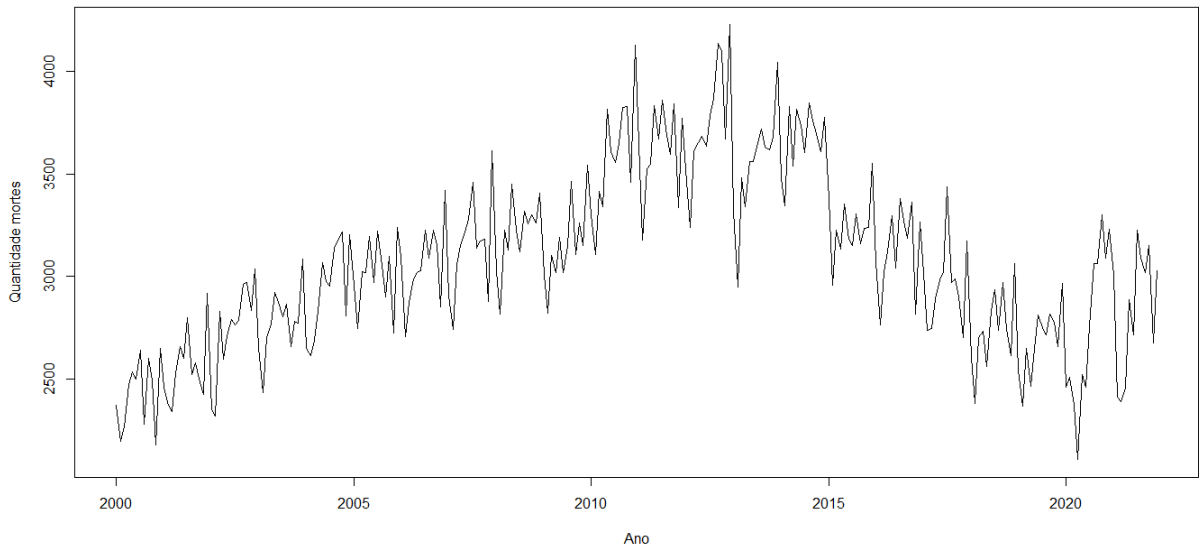

Figura 4.6: Mortalidade por acidente de trânsito no Brasil de janeiro de 2000 a dezembro de 2022.

Outra análise que pode nos auxiliar quanto a estacionariedade da série é o gráfico da Função de Autocorrelação (ACF). Caso essa função apresente uma convergência rápida para zero, pode-se concluir que a série é estacionária. No estudo em questão, essa função está apresentada na Figura 4.7, e não apresenta essa convergência, dando mais evidências de que a série é realmente não estacionária. Adicionalmente, pode-se observar que como todos os lags da ACF são significativos, tem-se indícios que a série possui tendência.

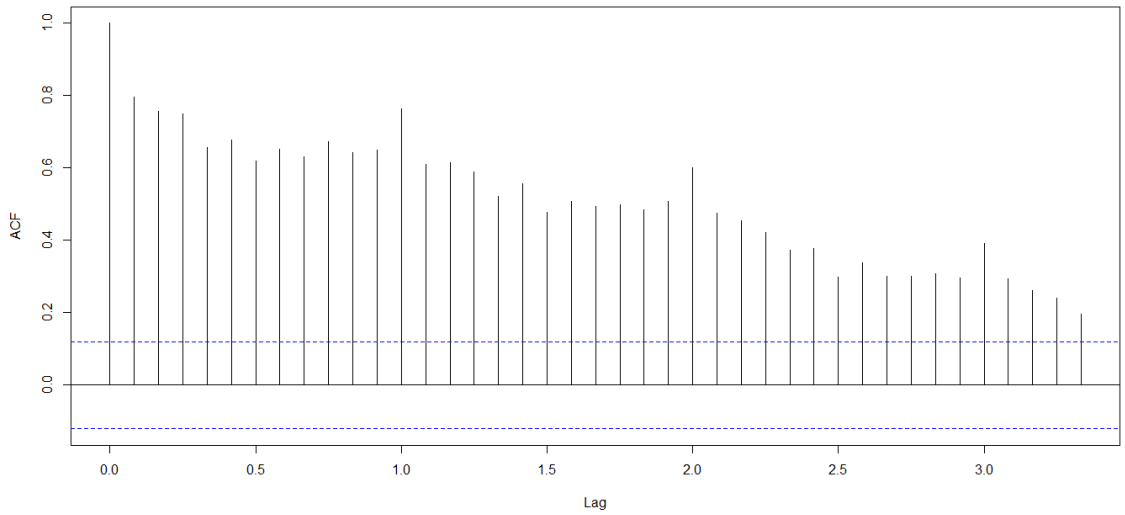

Figura 4.7: Função de Autocorrelação (ACF) da série temporal.

Para identificar uma possível sazonalidade, foi analisado o gráfico da Função de Autocorrelação Parcial (PACF), ilustrado na Figura 4.8, como ele possui lags significativos de multiplicidade 12, temos indícios da presença de sazonalidade estocástica na série.

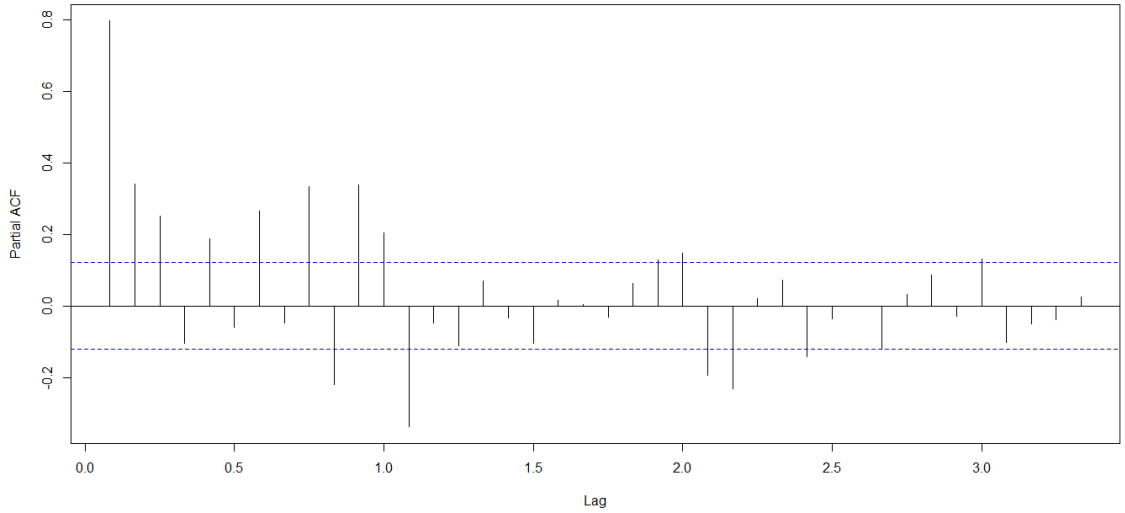

Figura 4.8: Função de Autocorrelação Parcial (PACF) da série temporal.

Um método utilizado para verificar a necessidade de transformação nos dados é analisar o gráfico de dispersão da média em função do desvio padrão. Caso ele apresente um comportamento aleatório não é necessário realizar nenhum tipo de transformação nos dados, pois isso indica que o desvio padrão é estável e não depende da média. Desse modo, a série original foi

dividida em grupos com 2 observações consecutivas, e para cada grupo foi calculada a média e o desvio padrão, cujos resultados foram dispostos em pares ordenados. Cada ponto no gráfico representa essas medidas, cujos resultados são apresentados na Figura 4.9. Analisando os pontos, pode-se notar que o desvio padrão não depende da média e, portanto, não é necessário transformar os dados.

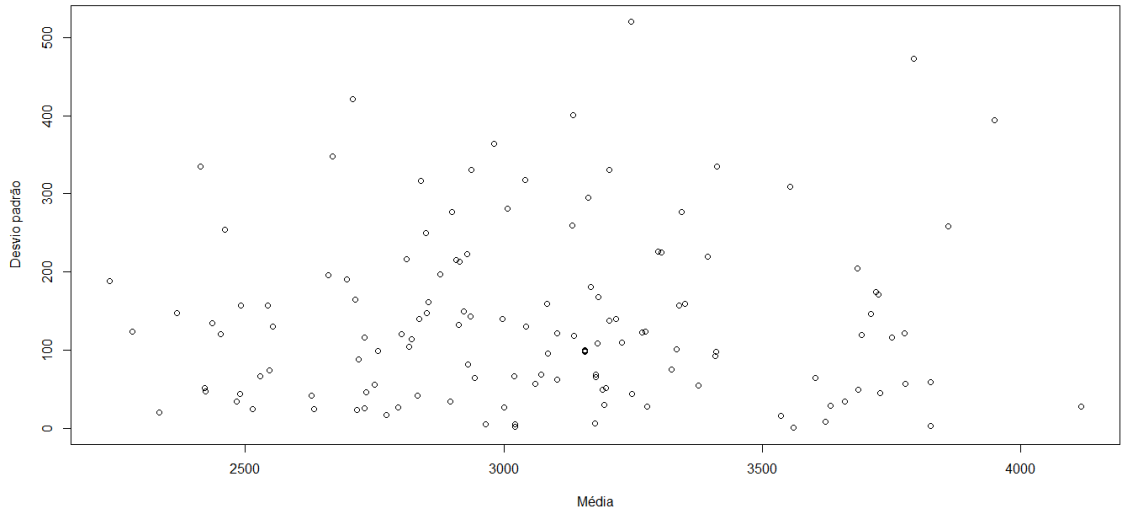

Figura 4.9: Cruzamento da média em relação ao desvio padrão da série original de Mortalidade por acidente de trânsito no Brasil.

O próximo passo foi realizar o teste de Mann-Kendall, afim de verificar a existência de tendência na série. Foi obtido um p-valor = 0,0001, resultado inferior a 5%, dessa forma a hipótese nula H_0 foi rejeitada, concluindo que a série possui tendência. Sendo assim, foi necessário aplicar uma diferença na série para remover a tendência. Na sequência, foi aplicado novamente o teste de Mann-Kendall para verificar se a tendência teria sido eliminada, em que foi obtido um p-valor = 0,8575, portanto, a hipótese nula H_0 não foi rejeitada, e concluiu-se que apenas uma diferenciação foi suficiente para eliminar a tendência da série. Nas Figuras 4.10 e 4.11 são apresentadas a série diferenciada e a ACF da série diferenciada, respectivamente.

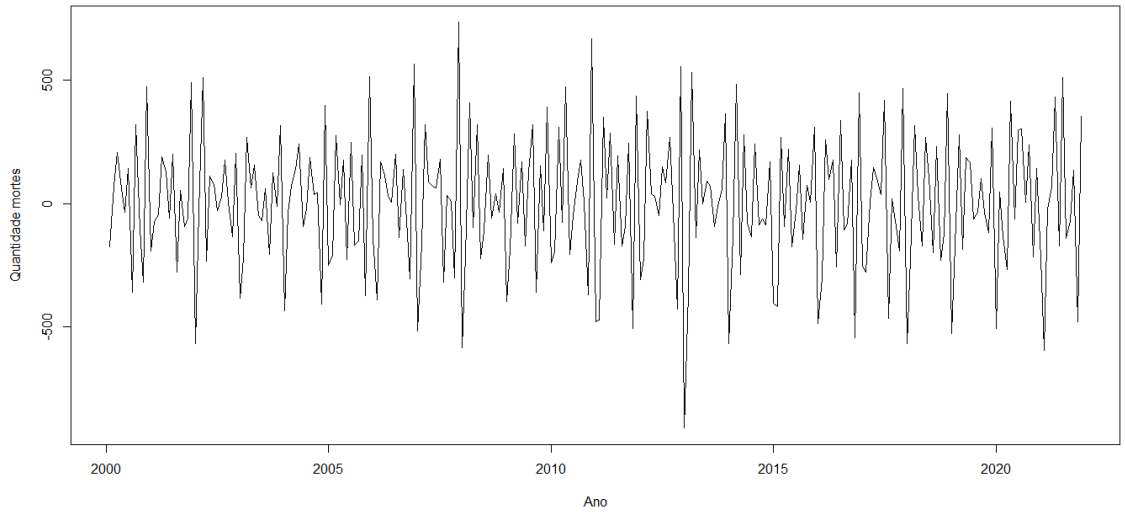

Figura 4.10: Série temporal diferenciada em grau 1.

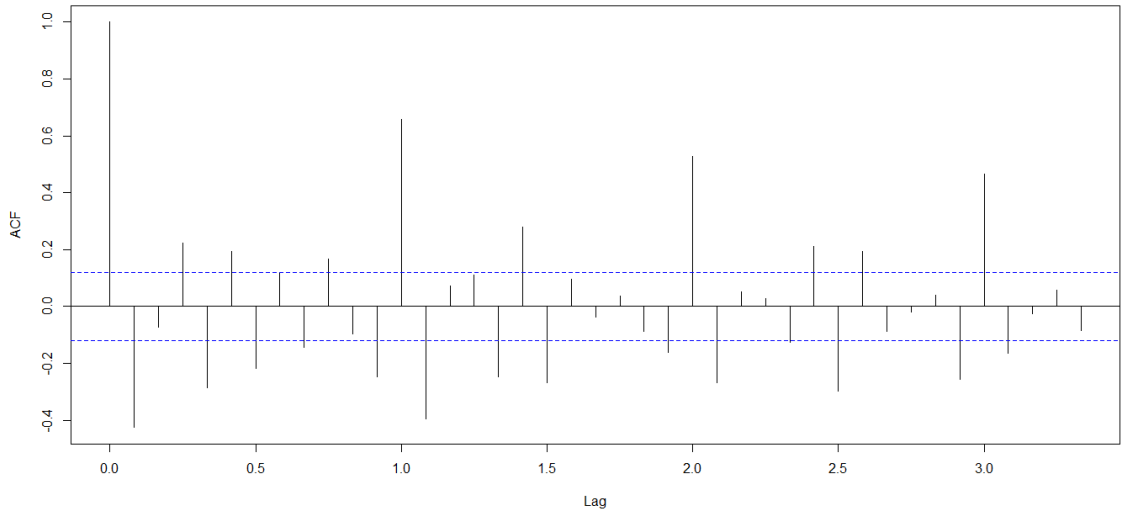

Figura 4.11: Função de Autocorrelação (ACF) da série temporal diferenciada.

Após remover a componente de tendência da série, foi realizado o teste de Fisher para sazonalidade. Foi obtido um p-valor = 1,46 x 10^{-57}, portanto a hipótese nula foi rejeitada, o que nos diz que a série possui sazonalidade. Esse resultado evidencia a necessidade de se ajustar modelos autorregressivos integrados de médias móveis sazonais (SARIMA).

A partir dessas premissas, o próximo passo foi encontrar o modelo que mais se adequa aos dados. Inicialmente, utilizando a função *auto.arima*, foi gerado um modelo automático considerado como ideal, e em seguida foram ajustados outros 7 modelos, sugeridos também por essa função *arima*, para que fossem comparados entre si e identificado o mais adequado entre eles, todos os modelos estão presentes na Tabela 4.1.

Modelo	Parâmetros	AIC	BIC
Modelo Indicado	SARIMA(5,1,1)(2,0,0)$_{12}$	3421,73	3453,88
Modelo 1	SARIMA(3,1,1)(1,0,0)$_{12}$	3459,85	3481,28
Modelo 2	SARIMA(1,1,0)(1,0,0)$_{12}$	3502,32	3513,04
Modelo 3	SARIMA(0,1,1)(0,0,1)$_{12}$	3547,99	3558,71
Modelo 4	SARIMA(2,1,1)(0,1,1)$_{12}$	3239,05	3256,68
Modelo 5	SARIMA(2,1,1)(2,0,0)$_{12}$	3429,55	3450,99
Modelo 6	SARIMA(2,1,1)(3,1,1)$_{12}$	3222,78	3250,98
Modelo 7	SARIMA(2,1,1)(3,1,0)$_{12}$	3246,26	3270,94

Tabela 4.1: Avaliação dos modelos SARIMA estudados.

Com exceção do modelo indicado, todos os outros modelos ajustados possuem os parâmetros significativos, então foram considerados como melhores os modelos que possuiam o menor Critério de Informação de Akaike (AIC) e o menor Critério de Informação Bayesiano (BIC), sendo estes os modelos 4, 6 e 7.

Para selecionar apenas um modelo, foi obtido o valor dos métodos residuais MAPE e RMSE utilizando os valores previstos de cada um desses modelos, que estão apresentados na Tabela

4.2.

Modelo	MAPE	RMSE
Modelo 4	3,57	142,39
Modelo 6	3,36	135,21
Modelo 7	3,64	144,51

Tabela 4.2: Avaliação dos métodos residuais MAPE e RMSE.

Pode-se observar na Tabela 4.2, que além de possuir o menor BIC e menor AIC, o modelo 6 possui os menores valores de MAPE e RMSE, portanto, ele foi escolhido como o ideal dentro desse estudo. Para validar a adequabilidade do modelo aos dados, foram analisadas graficamente as funções de autocorrelação (ACF) e autocorrelação parcial (PACF) dos resíduos, ilustrados nas Figuras 4.12 e 4.13, respectivamente.

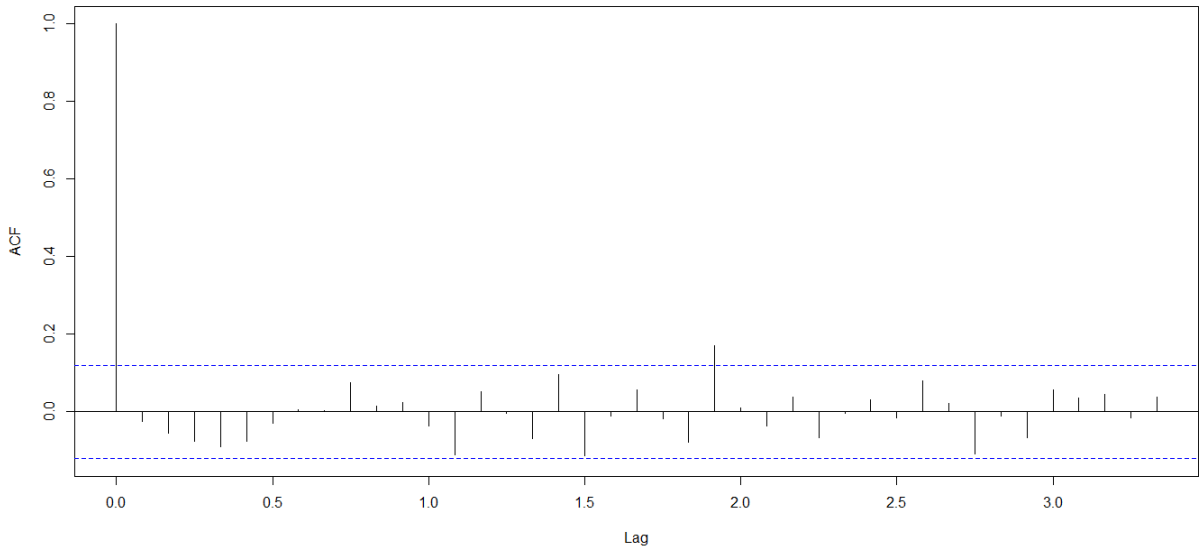

Figura 4.12: Função de Autocorrelação (ACF) dos resíduos.

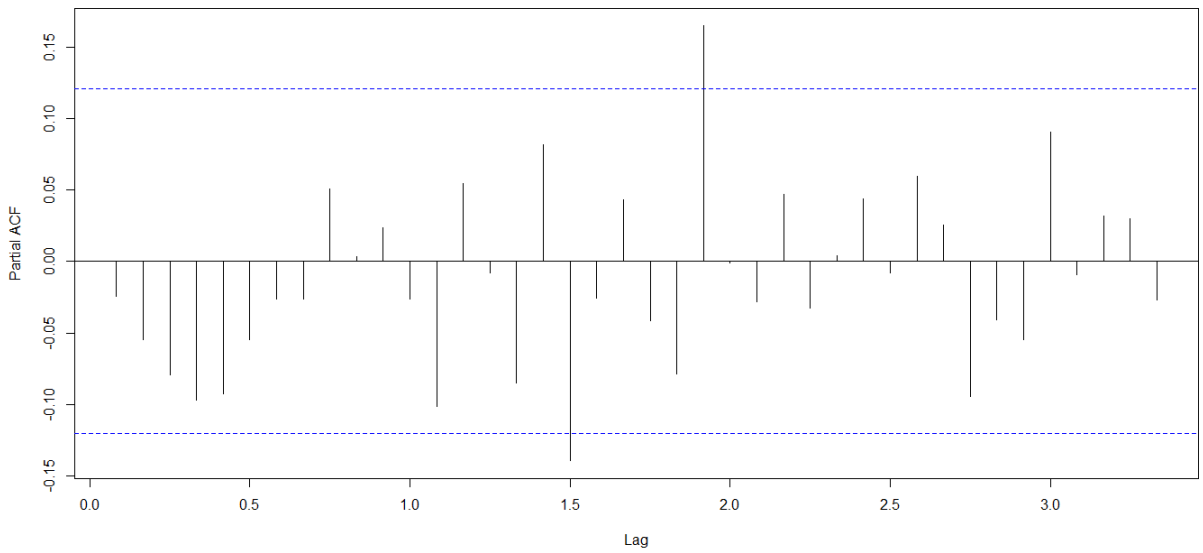

Figura 4.13: Função de Autocorrelação Parcial (PACF) dos resíduos.

Em ambos os gráficos existem no máximo 2 lags significativos, que seria o limite máximo aceitável, indicando que os resíduos não são correlacionados. Para confirmar essa não correlação, foi aplicado o teste de Box-Pierce sob a hipótese nula de que os resíduos são ruído branco. Foi obtido um p-valor = 0,6885, confirmando que os resíduos seguem um comportamento de ruído branco. De forma visual, foi plotado o periodograma acumulado dos resíduos, ilustrado na Figura 4.14, reforçando a conclusão obtida.

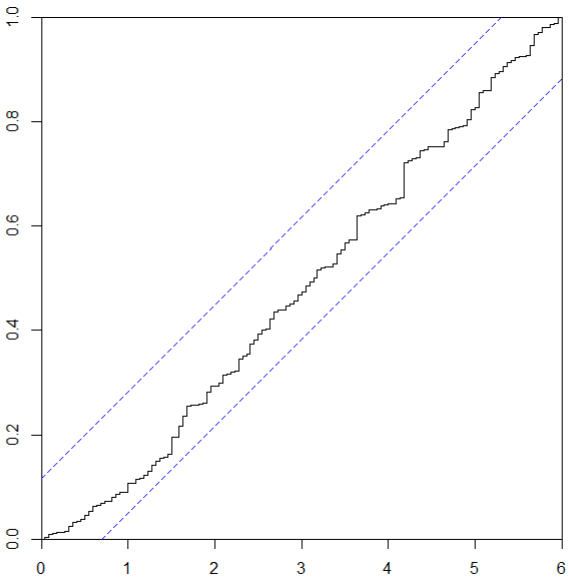

Figura 4.14: Peridiograma acumulado.

Por fim, agora com o modelo ajustado e validado, obtiveram-se as previsões para os meses de janeiro a dezembro de 2022, para serem comparados com os valores reais. Os valores observados e previstos de janeiro a dezembro de 2022, estão apresentados na Tabela 4.3. Na Figura 4.15 tem-se a ilustração gráfica desses resultados. De modo geral, as previsões se mantiveram próximas dos valores reais, chegando a uma variação máxima de 7,55% quando analisadas as previsões mensais, e quando analisadas as mortes totais do ano uma variação de 1,57%

Mês	Real	Previsto	Diferença
Jan/2022	2.486	2.640	154
Fev/2022	2.441	2.290	-151
Mar/2022	2.569	2.454	-115
Abr/2022	2.572	2.479	-93
Mai/2022	2.779	2.626	-153
Jun/2022	2.676	2.601	-75
Jul/2022	2.965	2.903	-62
Ago/2022	2.785	2.781	-4
Set/2022	2.792	2.800	8
Out/2022	2.987	2.801	-186
Nov/2022	2.567	2.530	-37
Dez/2022	2.717	2.922	205
Total	32.336	31.827	-509

Tabela 4.3: Resultado das previsão do modelo SARIMA(2,1,1)(3,1,1)12

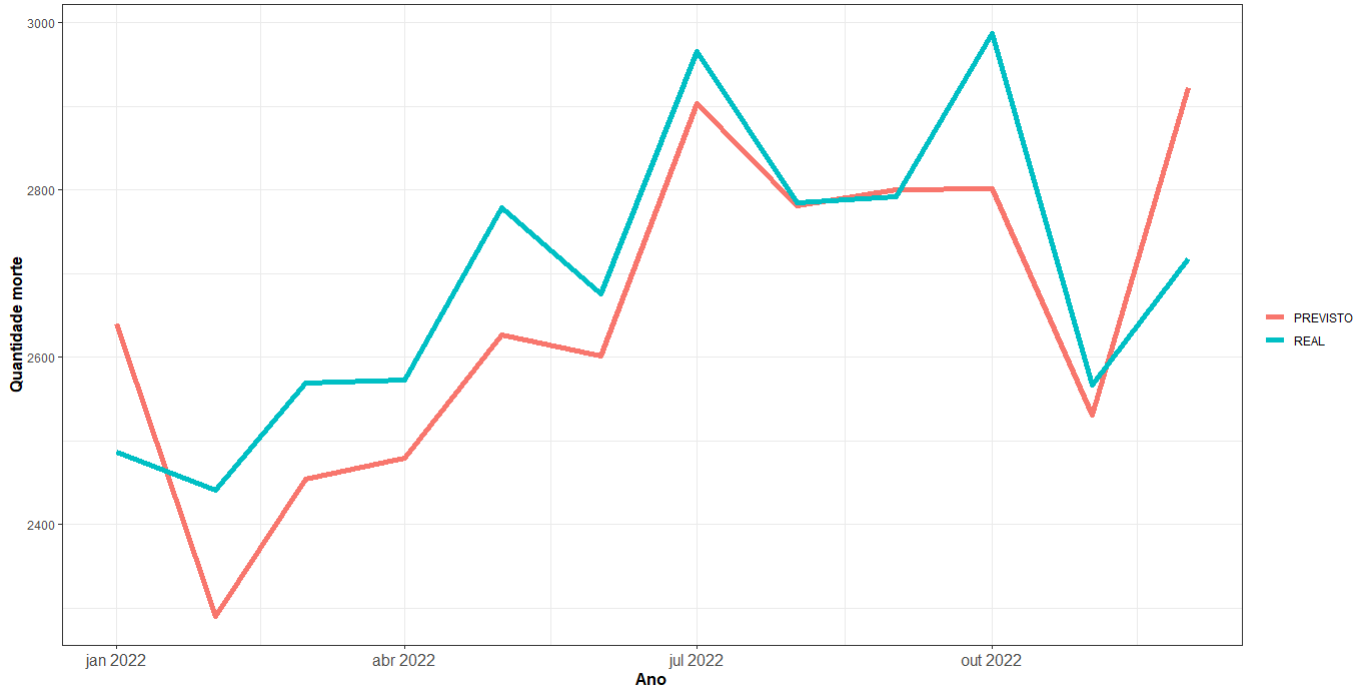

Figura 4.15: Comparativo da previsão com o valor real.

5. Conclusões

O modelo SARIMA(2,1,1)$(3,1,1)_{12}$, derivado da técnica de Box e Jenkins, que incorpora métodos autorregressivos sazonais de médias móveis, demonstrou valores reduzidos de MAPE e RMSE, revelando-se eficaz na compreensão e previsão da mortalidade por acidentes de trânsito no Brasil.

Com base nas previsões realizadas, é possível identificar que o período de final de ano é o momento em que esses eventos são mais frequentes, fornecendo assim informações importantes para a formulação de políticas públicas, como o aumento de blitz de trânsito nas rodovias e uma ampla campanha de conscientização sobre como conduzir com prudência e antecipar riscos, voltadas para a redução desses eventos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Ahdesmaki, M. Genecycle: Identification of periodically expressed genes, 2021. acessado em 27/11/2023. URL: https://cran.r-project.org/web/packages/GeneCycle/index.html.

[2] Akaike, H.A. *New Look at the Statistical Model Identification*. IEEE Transactions on Automatic Control, Minato-ku, v.19, n.6, p.716-723, 1974.

[3] Bertolo, L.A. Métodos básicos de previsão no excel, 2013. acessado em 07/04/2024. URL: http://www.bertolo.pro.br/MetodosQuantitativos/Simulacao/MetodosBasicosDePrevisaoDeSeriesTemporaisNoExcel.pdf.

[4] Carellos, L. Brasil é o terceiro país com maior número de mortes no trânsito, 2022. acessado em 27/11/2023. URL: https://www.vrum.com.br/noticias/mortes-transito-brasil-terceiro/.

[5] Fava, V.L. *Manual de econometria*. In: Vasconcelos, M. A. S.; Alves, D. São Paulo: Editora Atlas, 2000.

[6] Fokianos, K. and Strimmer, K. Fisher's exact g test for multiple (genetic) time series, 2004. acessado em 27/11/2023. URL: https://search.r-project.org/CRAN/refmans/GeneCycle/html/fisher.g.test.html.

[7] Governo Federal. Sistema de informação sobre mortalidade – sim, 2022. acessado em 27/11/2023. URL: https://dados.gov.br/dados/conjuntos-dados/sim-1979-2019.

[8] Hipel, K.W. and McLeod, A.I. *Time Series Modelling of Water Resources and Environmental Systems*. New York: Elsevier Science, 1994.

[9] Hyndman, R. Forecasting functions for time series and linear models, 2022. acessado em 27/11/2023. URL: https://pkg.robjhyndman.com/forecast/.

[10] Instituto de Pesquisa Econômica Aplicada. Estudo aponta aumento de 13,5 acessado em 27/11/2023. URL: https://www.ipea.gov.br/portal/categorias/45-todas-as-noticias/noticias/13899-estudo-aponta-aumento-de-13-5-em-mortes-no-transito.

[11] Morettin, P. A. e Toloi, C. M. C. *Análise de séries temporais*. Blucher, 2ª ed, 2006.

[12] Muianga, C.A. Descrição da curva de crescimento de frutos do cajueiro por modelos não lineares. 2016. URL: https://www.scielo.br/j/rbf/a/GTQ3pX9q38ZpZv47LQvzdbm/?lang=pt#.

[13] Paixão, L.M.M.M; Gontijo, E.D.; Mingoti, S.A.; Costa, D.A.S; Friche, A.A.L; Caiaffa, W.T. Óbitos no trânsito urbano: qualificação da informação e caracterização de grupos vulneráveis. 2014. URL: https://www.scielo.br/j/csp/a/4fCqZbygvgTjT3RPNytcL6R/?format=pdf&lang=pt.

[14] Pedersen, T.L. The grammar of graphics, 2022. acessado em 27/11/2023. URL: https://ggplot2.tidyverse.org/.

[15] Pohlert, T. Non-parametric trend tests and change-point detection, 2015. acessado em 27/11/2023. URL: https://www.rdocumentation.org/packages/trend/versions/0.0.1.

[16] Possato. Ajuste do arima para análise e previsões de séries temporais, 2020. acessado em 27/11/2023. URL: https://rpubs.com/Possato/ARIMA_ajuste_tutorial.

[17] Priestley, M. B. *Spectral Analysis and Time Series*. London: Academic Press. 407p, 1989.

[18] Reis, M.M. Análise de séries temporais, 2015. acessado em 07/04/2024. URL: http://www.inf.ufsc.br/~marcelo/Cap4.pdf.

[19] J.F. Santos. Tendências em séries de precipitação mensal em portugal continental. aplicação do teste de mann-kendall. pages 3–4, 2005. URL: https://comum.rcaap.pt/bitstream/10400.26/1303/1/Artigo_8SILUSBA_JFSantos_MMPortela_2007.pdf.

[20] Saragiotto, D. Acidentes de trânsito no brasil têm custo estimado de r$ 20 bilhões por ano, 2023. acessado em 27/11/2023. URL: https://mobilidade.estadao.com.br/mobilidade-para-que/acidentes-de-transito-no-brasil-tem-custo-estimado-de-r-20-bilhoes-por-ano/.

[21] Walter, F.C.; Maria, O.; Henning; Elisa; Moro; Graciela; Samohyl, W.; Robert. Aplicação de um modelo sarima na previsão de vendas de motocicletas. pages 77–88, 2013. URL: https://www.redalyc.org/pdf/810/81027458007.pdf.

[22] World Health Organization. Global status report on road safety - time for action, 2023. acessado em 27/11/2023. URL: https://www.afro.who.int/publications/global-status-report-road-safety-time-action.

A. DADOS

Ano/Mês	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro
2000	2369	2194	2264	2472	2532	2497	2640	2280	2600	2495	2177	2650
2001	2458	2385	2340	2531	2658	2598	2799	2522	2577	2482	2423	2915
2002	2348	2319	2830	2597	2710	2789	2761	2786	2961	2968	2834	3037
2003	2654	2432	2701	2766	2921	2872	2804	2863	2657	2782	2771	3086
2004	2650	2615	2687	2827	3067	2972	2951	3136	3173	3215	2808	3205
2005	2956	2747	3023	3020	3196	2971	3220	3052	2899	3097	2723	3238
2006	3095	2704	2872	2988	3018	3025	3227	3088	3226	3156	2851	3418
2007	2902	2740	3059	3148	3217	3279	3458	3138	3172	3181	2879	3614
2008	3028	2817	3227	3129	3450	3227	3118	3316	3258	3298	3262	3405
2009	3006	2819	3101	3020	3189	3017	3146	3465	3105	3259	3149	3540
2010	3301	3106	3415	3338	3813	3606	3556	3648	3823	3828	3459	4127
2011	3649	3175	3525	3548	3833	3669	3861	3689	3596	3842	3335	3772
2012	3463	3237	3612	3653	3684	3636	3784	3867	4136	4097	3670	4227
2013	3316	2949	3480	3341	3560	3559	3650	3720	3628	3616	3677	4042
2014	3475	3344	3828	3539	3817	3737	3604	3846	3760	3696	3608	3777
2015	3373	2955	3225	3132	3354	3180	3151	3306	3161	3234	3240	3550
2016	3065	2764	3023	3120	3296	3040	3378	3271	3187	3362	2816	3266
2017	3016	2738	2744	2891	2982	3020	3438	2971	2989	2897	2703	3170
2018	2603	2381	2699	2732	2561	2831	2936	2738	2968	2739	2615	3063
2019	2537	2367	2646	2462	2648	2813	2749	2713	2815	2777	2659	2965
2020	2459	2507	2372	2105	2522	2459	2756	3060	3064	3302	3087	3228
2021	3006	2411	2389	2456	2887	2716	3227	3087	3017	3152	2673	3027
2022	2486	2441	2569	2572	2779	2676	2965	2785	2792	2987	2567	2717

Tabela A.1: Série temporal da mortalidade por acidente de trânsito no Brasil, de Jan/2000 a Dez/2022.

B. Ajuste do Modelo SARIMA

Rotina implementada no R para o procedimento de Box e Jenkins, resultando em um modelo do tipo SARIMA:

```
# Série original
library(readxl)
dados <- read_excel("C:/Users/arthu/OneDrive/UFU/TCC/dados/DADOS V000-V999.xlsx", col_types = c("text", "numeric"))
dados <- dados[1:264,]
dados <- dados$MORTALIDADE
dados <- as.table(dados)
dados <- ts(dados,start=2000,frequency=12)
ts.plot(dados, xlab = "Ano", ylab = "Quantidade mortes",lwd = 3)

# FAC
acf(dados, lag.max = 40)

# FACp
pacf(dados, lag.max = 40)


# Calculando a média e o desvio media <- c()
j <- 1
vet <- dados
for (i in 1:nrow(dados)) {
media[i] <- (vet[j]+vet[j+1])/2
j <- j + 2
}
media
desvio <- c()
z <- 1
for (i in 1:nrow(dados)) {
desvio[i] <- sd(c(vet[z],vet[z+1]))
z <- z+2
```

```
    }
    desvio

# Verificar a necessidade de fazer transformção
    plot(media,desvio, xlab = "Média", ylab = "Desvio padrão")

# Teste de Mann-Kendall
    # Testar se apresenta tendência
    require(trend)
    mk.test(dados)

# Série diferenciada
    diffdados=diff(dados)
    ts.plot(diffdados)
    mk.test(diffdados)

# FAC série diferenciada
    acf(diffdados, lag.max = 40)

# FACp série diferenciada
    pacf(diffdados, lag.max = 40)
    ts.plot(diffdados, xlab = "Ano", ylab = "Quantidade mortes")

# Teste de sazonalidade - Pacote: GeneCycle
    # Testando se a série apresenta sazonalidade (H0 : NÃO tem sazonalidade)
    library(GeneCycle)
    fisher.g.test(dados)

# Ajuste dos modelos
    # Ajuste automático
    library(forecast)
    modelo <- auto.arima(dados)
    modelo
    modelo <- auto.arima(dados, trace = T)
    modelo

# Modelo Indicado
    # SARIMA(5,1,1)(2,0,0)12;
    M=arima(dados,order=c(5, 1, 1),seasonal=list(order=c(2, 0, 0)))
    library(lmtest)
```

```
    coeftest(M)
    BIC(M)  3453.88

# Modelo 1
    # SARIMA(3,1,1)(1,0,0)12;
    M1=arima(dados,order=c(3, 1, 1),seasonal=list(order=c(1, 0, 0)))
    coeftest(M1)
    BIC(M1)  3481.28
    M1

# Modelo 2
    # SARIMA(1,1,0)(1,0,0)12;
    M2=arima(dados,order=c(1, 1, 0),seasonal=list(order=c(1, 0, 0)))
    M2
    coeftest(M2)
    BIC(M2)  3513.04

# Modelo 3
    # SARIMA(0,1,1)(0,0,1)12;
    M3=arima(dados,order=c(0, 1, 1),seasonal=list(order=c(0, 0, 1)))
    M3
    coeftest(M3)
    BIC(M3)  3558.70

# Modelo 4
    # SARIMA(2,1,1)(0,1,1)12;
    M4=arima(dados,order=c(2, 1, 1),seasonal=list(order=c(0, 1, 1)))
    M4
    coeftest(M4)
    BIC(M4)  3256.68
    acf(M4$residuals, lag.max = 40)
    pacf(M4$residuals, lag.max = 40)

# Modelo 5
    # SARIMA(2,1,1)(2,0,0)12;
    M5=arima(dados,order=c(2, 1, 1),seasonal=list(order=c(2, 0, 0)))
    M5
    coeftest(M5)
    BIC(M5)  3450.99
    acf(M5$residuals, lag.max = 40)
```

```
    pacf(M5$residuals, lag.max = 40)

# Modelo 6
    # SARIMA(2,1,1)(3,1,1)12;
    M6=arima(dados,order=c(2, 1, 1),seasonal=list(order=c(3, 1, 1)))
    M6
    coeftest(M6)
    BIC(M6)  3250.98
    acf(M6$residuals, lag.max = 40)
    pacf(M6$residuals, lag.max = 40)

# Modelo 7
    # SARIMA(2,1,1)(3,1,0)12;
    M7=arima(dados,order=c(2, 1, 1),seasonal=list(order=c(3, 1, 0)))
    M7
    coeftest(M7)
    BIC(M7)  3270.94
    acf(M7$residuals, lag.max = 40)
    pacf(M7$residuals, lag.max = 40)

# Verificação de Normalidade dos Resíduos
    ts.plot(M6$residuals)
    hist(M6$residuals)

# Teste para verificar se os resíduos seguem a Normal (H0 : os resíduos seguem a Normal)
    Box.test(M6$residuals, type = c("Box-Pierce"))
    cpgram(M6$residuals) # Ruído Branco - Se meus resíduos seguem a normal

# Predição
    serie <- ts(dados, start = 2000, frequency = 12)
    M6=arima(dados,order=c(2, 1, 1),seasonal=list(order=c(3, 1, 1)))
    require(forecast)
    previsão <- forecast(M6, 12, level = c(95))
    accuracy(previsão)

# Carregando dados para plotar previsões
    library(readxl)
    previsões <- read_excel("C:/Users/arthu/OneDrive/UFU/TCC/dados/previsões.xlsx",sheet
= "Planilha2")
    str(previsões)
```

```
library(ggplot2)
ggplot(data = previsões, aes(x = Mês, y = VALOR, colour = SITUAÇÃO)) +
geom_line(size = 2)+
theme_bw()+
theme(axis.text.x = element_text(size = 12),
axis.title.x = element_text(size = 12,
face = "bold",
hjust = 0.5),
axis.title.y = element_text(size = 12,
face = "bold",
hjust = 0.5),
plot.title = element_text(size = 12,
face = "bold",
hjust = 0.5),
legend.title = element_blank())+
labs(x = "Ano",
y = "Quantidade morte")
```