



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

**Inflação no Brasil: uma aplicação de Séries
Temporais e Redes Neurais Recorrentes**

Julia Naves Rodrigues

Uberlândia-MG
2021

Julia Naves Rodrigues

**Inflação no Brasil: uma aplicação de Séries
Temporais e Redes Neurais Recorrentes**

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

Orientador: Prof^ª Dra. Nádia Giaretta Biase

Uberlândia-MG

2021



**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

Prof^a Dra. Nádia Giaretta Biase

Prof^a Dra. Gabriella de Freitas Alves

Prof^a Dra. Maria Imaculada de Sousa Silva

**Uberlândia-MG
2021**

AGRADECIMENTOS

Agradeço aos meus pais Éder e Giulliana, que sempre me mostraram que o estudo seria a ferramenta que transformaria minha vida, que nunca mediram esforços pra me proporcionar uma vida tranquila e uma educação de qualidade e que me incentivaram a correr atrás dos meus sonhos e sempre acreditaram neles também.

Ao meu irmão Pedro por sempre estar ao meu lado e me ouvir em meus momentos de tristeza e preocupação. Por ser meu parceiro nas madrugadas a dentro em meio aos estudos.

Ao meu marido Gilberto, que sempre se fez presente durante a graduação, e sempre me encorajou nas tomadas de decisões. Por estar sempre ao meu lado e me apoiar nas minhas decisões.

Agradeço também à professora e orientadora Nádia Giaretta Biase, que esteve me acompanhando desde minha entrada no Programa de Educação Tutorial (PET), que foi minha tutora durante 2 anos e pôde me acompanhar em uma Iniciação Científica. Por todos os conhecimentos que foram transmitidos, por suas correções, incentivos e conselhos.

À todos os amigos que fiz durante a graduação, que me proporcionaram momentos de alegria e acolhimento. Cada um tem um lugar especial na minha vida e vou levá-los para sempre.

Agradeço também às professoras Gabriella de Freitas Alves e Maria Imaculada de Sousa Silva, que apesar de não ter a oportunidade de conhecê-las pessoalmente, aceitaram prontamente o convite para participar da banca examinadora.

Por fim, agradeço aos meus familiares e amigos, que sempre me apoiam e torcem por minha felicidade e sucesso.

RESUMO

A realização de previsões da inflação é imprescindível quando se deseja melhorar o planejamento estratégico na intenção de reduzir incertezas e aumentar a capacidade de planejamento das famílias, empresas e governo, e assim garantir o bom funcionamento da economia do país. Por conta disso, o objetivo do presente estudo foi realizar previsões do Índice Nacional de Preços ao Consumidor Amplo (IPCA), o qual é a métrica oficial de mensuração da inflação pelo governo federal. As técnicas de séries temporais de Box e Jenkins e de redes neurais recorrentes foram utilizadas para construção de modelos que capturam as informações contidas na série, a fim de avaliar o comportamento da variável analisada ao longo do tempo, bem como efetuar previsões dos valores futuros. Ao aplicar as técnicas de Box e Jenkins, foi adotado um modelo de previsão do tipo SARIMA, que considera tanto a tendência como a sazonalidade na série, sendo selecionado aquele com menor raiz do erro quadrático médio (RMSE do inglês Root Mean Squared Error) e menor quantidade de parâmetros. Para os modelos de redes neurais recorrentes, utilizou-se o algoritmo LSTM (Long Short Term Memory), e selecionou-se aquele com menor RMSE. A fim de comparar as duas técnicas de previsão, os critérios adotados foram raiz do erro quadrático médio (RMSE), sendo que o modelo SARIMA se mostrou o melhor modelo para prever 12 meses diretos da inflação, enquanto que para previsões a curto prazo com realimentação, a técnica LSTM se mostrou muito efetiva.

Palavras-chave: IPCA, Política Monetária, Box & Jenkins, SARIMA, LSTM.

ABSTRACT

The realization of inflation forecasts is essential when one wants to improve strategic planning in order to reduce uncertainties and increase the planning capacity of families, companies and government, and thus ensure the proper functioning of the country's economy. The aim of this study was to make forecasts of the Broad National Consumer Price Index (IPCA), which is the official metric for measuring inflation by the federal government. The Box and Jenkins time series and recurrent neural network techniques were used to build models that capture the information contained in the series, in order to evaluate the behavior of the analyzed variable over time, as well as to make predictions of future values. By applying the Box and Jenkins techniques, a SARIMA-type forecasting model was adopted, which considers both the trend and the seasonality in the series, being selected the one with the smallest root mean squared error (RMSE from English RootMean Squared Error) and the smallest amount of parameters. For the models of recurrent neural networks, the LSTM (Long Short Term Memory) algorithm was used, and the one with the lowest RMSE was selected. In order to compare the two forecasting techniques, the adopted criteria were root mean square error (RMSE), and the SARIMA model proved to be the best model to predict 12 direct months of inflation, while for short-term forecasts with feedback, LSTM technique proved to be very effective.

Keywords: IPCA, Monetary Policy, Box & Jenkins, SARIMA, LSTM .

SUMÁRIO

Lista de Figuras	I
Lista de Tabelas	III
1 Introdução	1
2 Fundamentação Teórica	3
2.1 Séries Temporais	3
2.1.1 Modelos para Séries Temporais	3
2.1.2 Decomposição Clássica da Série	5
2.1.3 Modelos Sazonais	9
2.1.4 Critérios de seleção de modelos	10
2.1.5 Testes de Adequação do Modelo	11
2.2 Redes Neurais Artificiais (RNA)	12
2.2.1 Redes Neurais do tipo LSTM	13
3 Metodologia	17
4 Resultados	21
4.1 Cenário A	21
4.1.1 Modelos SARIMA	21
4.1.2 Modelos LSTM	27
4.1.3 Comparação entre os Modelos SARIMA e LSTM	31
4.2 Cenário B	32
4.2.1 Modelos SARIMA	32
4.2.2 Modelos LSTM	38
4.2.3 Comparação entre os Modelos SARIMA e LSTM	41
5 Conclusões	43
Referências Bibliográficas	45

LISTA DE FIGURAS

2.1	Estrutura de uma Rede Neural Artificial	12
2.2	Estrutura de uma Rede Neural Recorrente [27]	13
2.3	Estrutura de uma Rede LSTM	14
4.1	IPCA acumulado mensal de Dezembro/1999 à Abril/2021	21
4.2	Função de Autocorrelação da série (Cenário A)	22
4.3	Distribuição Média pela Amplitude de grupos de observações da série (Cenário A)	22
4.4	Série original com uma diferença (Cenário A)	23
4.5	FAC e FACp da série diferenciada (Cenário A)	24
4.6	Gráficos de FAC e FACp dos Resíduos (Cenário A)	26
4.7	Periodograma Acumulado (Cenário A)	26
4.8	Previsões do modelo SARIMA (Cenário A)	27
4.9	Previsões do modelo LSTM Sem Reposição (Cenário A)	29
4.10	Previsões do modelo LSTM Com Reposição (Cenário A)	30
4.11	Série Original e Previsões dos modelos (Cenário A)	31
4.12	IPCA acumulado mensal de Dezembro/1999 à Dezembro/2019	32
4.13	Função de Autocorrelação da série (Cenário B)	33
4.14	Distribuição Média pela Amplitude de grupos de observações da série (Cenário B)	33
4.15	Gráfico da série original com uma diferença (Cenário B)	34
4.16	Gráficos de FAC e FACp da série diferenciada (Cenário B)	35
4.17	Gráficos de FAC e FACp dos Resíduos (Cenário B)	36
4.18	Periodograma Acumulado (Cenário B)	37
4.19	Previsões do modelo SARIMA (Cenário B)	37
4.20	Previsões do modelo LSTM Sem Reposição (Cenário B)	39
4.21	Previsões do modelo LSTM Com Reposição (Cenário B)	40
4.22	Série Original e Previsões dos modelos (Cenário B)	41

LISTA DE TABELAS

4.1	Teste para verificação de tendência na série (Cenário A)	23
4.2	Teste para verificação de sazonalidade na série (Cenário A)	24
4.3	Ajustes de modelos SARIMA propostos	24
4.4	Critérios de seleção dos modelos SARIMA propostos (Cenário A)	25
4.5	Parâmetros estimados do modelo SARIMA $(1,1,0)(1,0,0)_{[12]}$	25
4.6	Previsão realizada utilizando modelo SARIMA e erro de previsão (Cenário A). .	27
4.7	Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Sem Reposição (Cenário A)	28
4.8	Previsão realizada utilizando modelo LSTM Sem Reposição (Cenário A)	29
4.9	Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Com Reposição (Cenário A)	30
4.10	Previsão realizada utilizando modelo LSTM Com Reposição e erro de previsão (Cenário A)	30
4.12	Valores de RMSE dos modelos selecionados de cada técnica estudada (Cenário A)	31
4.13	Teste para verificação de tendência na série (Cenário B)	34
4.14	Teste para verificação de sazonalidade na série (Cenário B)	35
4.15	Critérios de seleção dos modelos SARIMA propostos (Cenário B)	35
4.16	Parâmetros estimados do modelo SARIMA $(1,1,0)(2,0,0)_{[12]}$	36
4.17	Previsão realizada utilizando modelo SARIMA e erro de previsão (Cenário B) .	38
4.18	Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Sem Reposição (Cenário B)	38
4.19	Previsão realizada utilizando modelo LSTM Sem Reposição e erro de previsão (Cenário B)	39
4.20	Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Com Reposição (Cenário B)	40
4.21	Previsão realizada utilizando modelo LSTM Com Reposição e erro de previsão (Cenário B)	41
4.22	Valores de RMSE dos modelos selecionados de cada técnica estudada (Cenário B)	41

1. INTRODUÇÃO

Inflação é o nome dado ao aumento dos preços de bens e serviços, e implica na diminuição do poder de compra da moeda. Ela é calculada pelos índices de preços, comumente chamados de índices de inflação. Dois dos índices mais comumente utilizados são produzidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE): o Índice Nacional de Preços ao Consumidor (INPC) e o Índice Nacional de Preços ao Consumidor Amplo (IPCA)[18]. São vários os fatores que podem levar ao aumento deste fenômeno, entre eles as pressões de demanda e de custos, a inércia inflacionária e as expectativas de inflação. O impacto da inflação sobre a economia é grande, visto que gera incertezas, desestimulando o investimento e, assim, prejudicando o crescimento econômico.

O IPCA é o principal indicador utilizado para medir a inflação, sendo este considerado o oficial pelo governo federal. Seu objetivo é mensurar a variação de preços de uma cesta de produtos e serviços consumida pela população. O resultado mostra se os preços aumentaram ou diminuíram de um mês para o outro[18].

Manter a inflação sob controle, ao redor da meta, é objetivo fundamental do Banco Central (BACEN). A meta para a inflação é estabelecida pelo Conselho Monetário Nacional (CMN) e as taxas de juros são controladas pelo Comitê de Política Monetária do Banco Central. O BACEN tem como principal objetivo assegurar a estabilidade do poder de compra da moeda e um sistema financeiro sólido e eficiente. A instituição é responsável por executar a estratégia estabelecida pelo CMN para manter a inflação sob controle[5].

O cenário ideal para a inflação é de valores baixos, estáveis e previsíveis, trazendo vários benefícios para a sociedade, visto que a incerteza na economia é menor, as pessoas podem planejar melhor seu futuro e as famílias não têm sua renda real corroída. Por conta disso, o Brasil adota um sistema de metas para inflação, desde 1999, a fim de controlar esse índice. A meta se refere à inflação acumulada no ano e é definida anualmente no mês de junho, para três anos-calendário à frente, na intenção de reduzir incertezas e melhorar a capacidade de planejamento. O sistema ainda prevê um intervalo de 1,5 pontos percentuais, ou seja, no caso de 2020 que a meta é 4,00%, o intervalo é de 2,5% à 5,5%[5]. Caso a inflação final estiver fora do intervalo de tolerância, o presidente do BACEN deve divulgar publicamente as razões do descumprimento[4].

O Banco Central utiliza a política monetária para alcançar o objetivo de estabilizar os preços, e por consequência preservar o valor do dinheiro, mantendo o poder de compra da moeda. Essa política se refere às ações que visam afetar o custo do dinheiro (taxas de juros)

e a quantidade de dinheiro (condições de liquidez) na economia. O principal instrumento utilizado pelo BACEN é a taxa Selic, que por sua vez é definida pelo Comitê de Política Monetária do Banco Central (COPOM) a cada 45 dias[20]. Ela se refere à taxa apurada nas operações de empréstimos de um dia entre as instituições financeiras que utilizam títulos públicos federais como garantia. Ela influencia todas as taxas de juros do país, como juros de empréstimos, financiamentos e aplicações financeiras, além de operar por vários canais que acabam por influenciar o comportamento da inflação.

Tendo em vista esse cenário, é substancial a realização de previsões da inflação, a fim de melhorar o planejamento estratégico de famílias, empresas e governo e assim garantir o bom funcionamento da economia do país. Tradicionalmente, existem modelos de acordo com a teoria econômica que são utilizados para realizar previsões, que relacionam alguns índices na influência da variação dos preços de uma economia, sendo eles: taxa de desemprego, taxa de câmbio, taxa de juros e expectativas da inflação. Um exemplo de modelo tradicional é a curva de Phillips[6]. Além disso, existem diversas técnicas estatísticas e computacionais capazes de fazer essas estimativas, como por exemplo as técnicas propostas por Box & Jenkins e as técnicas de redes neurais artificiais.

Nas previsões modeladas através de séries temporais, a predição tem por base o estudo dos dados históricos de uma determinada variável de interesse ao longo de um período de tempo. Uma das principais vantagens das análises de modelos de Box & Jenkins se dá por conta de terem inúmeros recursos que podem ser levados em consideração com grande potencial para seleção de um método com melhor desempenho. Bonno (2014) desenvolveu um estudo aplicando modelos sazonais no IPCA e obteve resultados satisfatórios de predição[7]. Barbosa (2019) também propôs uma análise de previsão da inflação, mas desta vez utilizando modelos ARIMA, desconsiderando a componente sazonal, e comparou os resultados com um modelo de acordo com a teoria econômica[6].

Com a constante evolução de métodos computacionais e das técnicas de Inteligência Artificial, teve-se como resultado métodos de Redes Neurais Artificiais (RNA) capazes de prever comportamentos lineares e não lineares. As RNA's têm apresentado excelentes resultados nas previsões de séries temporais, principalmente quando se envolve grandes massas de dados, como em previsões do movimento de ações nas bolsas de valores, por exemplo[12].

Neste contexto, o presente estudo tem como objetivo a modelagem e estimação de valores do índice inflação no Brasil, medido pelo IPCA, através de métodos de séries temporais de Box & Jenkins e modelos de redes neurais artificiais, e dentre estes avaliar os que apresentarem melhores resultados de previsão.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 SÉRIES TEMPORAIS

Uma série temporal é um conjunto de dados observados sucessivamente ao longo do tempo, em intervalos uniformes. Esses dados são dependentes entre observações vizinhas e o objetivo do estudo é analisar e modelar essa dependência [21]. Além disso, vale notar que o tempo pode ser substituído por outra variável como espaço, profundidade, etc [22].

A série temporal pode ser classificada de duas formas: contínua ou discreta. Na primeira, os dados são observados de forma contínua ao longo do tempo e o conjunto T é definido por $T = \{t : t_1 < t < t_2\}$, enquanto que na segunda, os dados são feitos em tempos específicos equidistantes e o conjunto é definido por $T = \{t_1, \dots, t_n\}$. Esse tipo de classificação não diz respeito ao tipo de dado da variável observada, mas sim à forma como são coletados os dados. As séries contínuas podem ser discretizadas, de forma a registrar seus valores a certos intervalos de tempo. Séries de valores agregados ou acumulados em intervalos de tempo, por exemplo exportações medidas mensalmente ou quantidade de chuva medida diariamente [22].

Como a maior parte dos procedimentos estatísticos foram desenvolvidos para análise de observações independentes, o estudo de séries temporais exige o uso de técnicas específicas para cada área de aplicação, uma vez que esses dados surgem de diversos campos de conhecimento. Como exemplo, podemos citar a economia (preços diários de ações, taxa mensal de desemprego, inflação), medicina (eletrocardiograma, eletroencefalograma), epidemiologia (número mensal de novos casos de COVID-19, dengue e tipos de câncer), meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), etc [22].

Em suma, os principais objetivos deste tipo de estudo são: descrever padrões de tendência, ciclos e sazonalidade, identificar observações discrepantes (outliers), mudanças de padrões, explicar a variação de outra série relacionada, prever valores futuros e auxiliar na tomada de decisão [21] [22].

2.1.1 MODELOS PARA SÉRIES TEMPORAIS

Segundo Souza (1981)[28], podemos definir uma série temporal como uma realização qualquer de um processo estocástico. Seja T um conjunto arbitrário, uma família de variáveis aleatórias $Z = \{Z(t), t \in T\}$, tal que, para cada $t \in T$, $Z(t)$ é uma variável aleatória, diz-se que Z é um processo estocástico[21].

Em outras palavras, um processo estocástico é um conjunto de variáveis definido num espaço de probabilidades $(\Omega, \mathcal{A}, \mathcal{P})$. Normalmente, T é tomado como o conjunto dos inteiros Z , e o processo diz-se com parâmetro discreto, ou o conjunto dos reais R , em que obtemos um processo com parâmetro contínuo[21].

De acordo com Morettin e Toloí (2006)[21], quando se planeja empregar modelos para descrever séries temporais, faz-se necessário elencar as seguintes suposições simplificadoras, para que se permita analisar determinadas classes de processos estocásticos. Sendo assim:

- processos estacionários ou não estacionários, de acordo com a independência ou não relativamente à origem dos tempos;
- processos normais (Gaussianos) ou não normais, conforme as funções de distribuição de probabilidade que caracterizam os processos;
- processos Markovianos ou não-Markovianos, segundo a independência dos valores do processo, em dado instante, de seus valores em instantes precedentes.

Isto é, um processo Z é dito estacionário quando ele distender ao longo do tempo de forma que a escolha da origem dos tempos não seja relevante, isto é, as características de $Z(t+\tau)$, para todo τ , são as mesmas de $Z(t)$. Existem duas formas de estacionariedade: fraca (ou ampla) e estrita (ou forte), e serão definidas adiante.

Um processo estocástico $Z = \{Z(t), t \in T\}$ é dito *estritamente estacionário* quando todas as distribuições finito-espaciais permanecem as mesmas sob translações no tempo. Assim:

$$F(z_1, \dots, z_n; t_1 + \tau, \dots, t_n + \tau) = F(z_1, \dots, z_n; t_1, \dots, t_n), \forall t_1, \dots, t_n, \tau \text{ de } T. \quad (2.1)$$

Diante disso, todas as distribuições unidimensionais são invariantes sob translações no tempo, resultando em uma média $\mu(t) = \mu$ e em uma variância $V(t) = \sigma^2$ constantes para todo $t \in T$.

Por sua vez, um processo estocástico é definido como *fracamente estacionário* se, e somente se:

- (i) $E\{Z(t)\} = \mu(t) = \mu$, constante, $\forall t \in T$;
- (ii) $E\{Z^2(t)\} < \infty$, $\forall t \in T$;
- (iii) $\gamma(t_1, t_2) = Cov\{Z(t_1), Z(t_2)\}$ é uma função de $|t_1 - t_2|$.

Um exemplo de um processo estocástico é o ruído branco, que pode ser definido como uma sequência de variáveis randômicas independentes e identicamente distribuídas (i.i.d.), que seguem uma distribuição normal de média zero, variância constante e as covariâncias são nulas.

O presente trabalho está interessado em estudar apenas essa classe de processos, que serão denotados por processos estacionários.

2.1.2 DECOMPOSIÇÃO CLÁSSICA DA SÉRIE

Tendo visto sobre a estacionariedade, faz-se necessário o conhecimento sobre tendência e sazonalidade, uma vez que esses conceitos são essenciais na construção do modelo de séries temporais.

O modelo de decomposição de uma série temporal pode ser escrito pela soma ou produto de três componentes não observáveis: T_t que indica tendência, S_t que indica sazonalidade e a_t que é uma componente aleatória que segue uma distribuição normal de média zero e variância constante, também chamada de erro.

$$Z_t = T_t + S_t + a_t \quad (2.2)$$

$$Z_t = T_t * S_t * a_t \quad (2.3)$$

Segundo Kopecky e Pierce (1979)[19], as componentes T_t e S_t são, em geral, bastante relacionadas e a influência da tendência sobre a componente sazonal pode ser muito forte, por duas razões:

1. métodos de estimação de S_t podem ser bastante afetados se não levarmos em conta a tendência;
2. a especificação de S_t depende da especificação de T_t

Por conta destes fatos, não podemos isolar uma das componentes sem tentar isolar a outra. Estimando T_t e S_t e subtraindo de Z_t encontramos uma estimativa da componente aleatória a_t .

Modelos aditivos são mais adequados quando a sazonalidade não depende das outras componentes. Já quando a componente sazonal varia com a tendência, o modelo multiplicativo é indicado[21].

TENDÊNCIA

De acordo com Ehlers (2009)[22], podemos pensar em tendência como uma mudança de longo prazo no nível médio da série. A dificuldade aqui é definir longo prazo.

Conforme Morettin e Tolo (2006)[21], há vários métodos para estimação de T_t . Os mais comuns são:

- ajuste de uma função polinomial;
- suavização dos valores da série ao redor de um ponto, para estimar a tendência naquele ponto (médias móveis, medianas móveis);
- sucessivos ajustes de retas de mínimos quadrados ponderados (“lowess”).

Inicialmente para identificar a tendência, pode-se utilizar uma regressão linear simples para avaliar graficamente o comportamento dos dados, de tal modo que seja possível observar se existe uma tendência positiva ou negativa. Esse tipo de análise pode ser complementado com o uso de testes de hipóteses estatísticos, que podem ser feitos depois da estimação de T_t ou, preferivelmente, antes[3].

Existem testes não paramétricos específicos que são bastante utilizados para a identificação da tendência no conjunto de dados observados. Morettin e Tolo (2006)[21] citam os seguintes testes: teste de Wald-Wolfowitz e o teste de Cox-Stuart, e para ambos avalia-se a aceitação da hipótese nula $H_0 = \text{Não existe tendência}$.

O teste de Wald-Wolfowitz (ou teste de sequências), compara cada observação com a mediana e atribui o símbolo A caso a observação seja maior ou igual à mediana, e B caso seja menor. A partir disto, define-se N_a como a quantidade de pontos atribuídos com A , N_b a quantidade de pontos atribuídos com B e $N = N_a + N_b$ a quantidade total de observações da série. A estatística T_1 utilizada no teste será dada pelo número total de grupos com símbolos iguais e comparada à estatística tabelada w_α , que é o quantil da distribuição de T_1 , irá indicar a rejeição ou não da hipótese nula. Para uma série com mais de 20 observações, é possível utilizar uma aproximação normal, isto é, $T_1 \sim N(\mu, \sigma^2)$ onde:

$$\mu = \frac{2N_a N_b}{N} + 1; \quad (2.4)$$

$$\sigma^2 = \frac{2N_a N_b (2N_a N_b - N)}{N^3 - N} \quad (2.5)$$

Para o teste de Cox-Stuart (ou teste do sinal), são agrupadas as observações em pares $(Z_1, Z_{1+c}), (Z_2, Z_{2+c}), \dots, (Z_N, Z_{N+c})$ onde $c = N/2$ se N for par, e $c = (N + 1)/2$ se N for ímpar, sendo N o número total de observações. A cada par (Z_i, Z_{i+c}) é associado o sinal “+” se $(Z_i < Z_{i+c})$ e o sinal “-” se $(Z_i > Z_{i+c})$, eliminando os empates. Seja N_t a quantidade de pares em que $(Z_i \neq Z_{i+c})$, a estatística T_2 é o número de pares com sinal “+”, e para $T_2 > N_t - T_{Tab}$ rejeita-se a hipótese H_0 . O valor de T_{Tab} será definido de acordo com a quantidade de observações: se menor ou igual a 20 a regra de decisão será baseada na distribuição binomial, e se maior pode-se utilizar a aproximação normal.

Além destes testes, é possível testar a tendência da série por meio do coeficiente de correlação de Spearman, que em geral é mais poderoso que o teste de Cox-Stuart. Seja R_t o posto de Z_t dentre as N observações. Se houver observações empatadas, use postos médios. A estatística do teste T_3 é dada por:

$$T_3 = \sum_{t=1}^N [R_t - t]^2 \quad (2.6)$$

em que $t = 1, 2, \dots$, são pontos “naturais”: dos instantes de tempo. A estatística T_3 pode ser comparada com o valor tabelado da distribuição T_3 , que pode ser encontrada em (Conover,

1980)[11], ou com a função do coeficiente de correlação de Spearman, dado por[21]:

$$\rho = 1 - \frac{6T_3}{N^3 - N} \quad (2.7)$$

A distribuição do coeficiente de Spearman pode ser aproximada de uma distribuição T de Student, com $n - 2$ graus de liberdade. Para verificar a hipótese nula de independência, neste caso, ausência de tendência, é possível aplicar um teste dado por:

$$t = \sqrt{\frac{(n - 2)\rho^2}{(1 - \rho^2)}} \quad (2.8)$$

Identificada a existência de tendência e estimada por \hat{T}_t , é possível obter a série ajustada, dada por:

$$Y_t = Z_t - \hat{T}_t \quad (2.9)$$

Pode-se também eliminar a tendência de uma série tomando uma ou duas diferenças sucessivas. A primeira, pode ser expressa como:

$$\Delta Z_t = Z_t - Z_{t-1}. \quad (2.10)$$

SAZONALIDADE

Entende-se por sazonalidade comportamentos que ocorrem regularmente dentro de um curto período de tempo, geralmente de um ano para outro, como por exemplo o aumento de venda de passagens aéreas no verão, aumento da produção de leite no Brasil nos meses de novembro, dezembro e janeiro, e aumento de venda do comércio na época do Natal[21].

No estudo de séries temporais, o procedimento de ajuste sazonal consiste em:

(i) obter as estimativas \hat{S}_t de S_t ;

(ii) calcular

$$Z_t^{SA} = Z_t - \hat{S}_t \quad (2.11)$$

para modelos aditivos, ou

$$Z_t^{SA} = \frac{Z_t}{\hat{S}_t} \quad (2.12)$$

para modelos multiplicativos.

Morettin e Toloi (2006)[21] propõem que existem dois tipos de sazonalidade: a determinística e a estocástica. A sazonalidade determinística ocorre quando o padrão sazonal é regular e estável no tempo e o comportamento sazonal pode ser previsto a partir de dados anteriores. Neste caso, a estimação por meio de método de regressão é ideal. Por outro lado, quando a componente sazonal varia com o tempo, trata-se de uma sazonalidade estocástica, sendo o método adequado para estimação de \hat{S}_t o de médias móveis.

Para verificar a sazonalidade determinística, alguns testes não paramétricos podem ser utilizados. O primeiro teste proposto por Morettin e Toloi (2006)[21] é o teste de Kruskal-Wallis. Neste, dispõe-se em tabela as observações em k (colunas) de tamanho n_j (linhas), e cada k é considerado uma amostra da população. As observações são denotadas por:

$$Y_{ij}, j = 1, \dots, k, i = 1, \dots, n_j, N = \sum_{j=1}^k n_j \quad (2.13)$$

As observações Y_{ij} são substituídas pelos postos R_{ij} , que por sua vez são obtidos ordenando todas as N observações. Seja $R_{.j}$ a soma dos postos associados à j -ésima coluna:

$$R_{.j} = \sum_{i=1}^{n_j} R_{ij}, j = 1, \dots, k. \quad (2.14)$$

Rejeita-se a hipótese H_0 de não existência de sazonalidade se a estatística do teste T_4 :

$$T_4 = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_{.j}^2}{n_j} - 3(N+1) \quad (2.15)$$

for maior ou igual ao valor crítico T_{4c} , onde T_{4c} é tal que $P_H(T_4 \geq T_{4c}) = \alpha$, e α o nível de significância do teste. Para n_j suficientemente grande, ou $k \geq 4$, sob H_0 , a distribuição de T_4 pode ser aproximada por uma variável χ^2 com $k - 1$ graus de liberdade. Uma crítica ao teste é que existem suposições de que existe independência entre as variáveis de uma amostra e as amostras possuem independência entre si, e no caso de séries temporais essas suposições não são verificadas[21].

Uma alternativa é o teste de Friedman para amostras relacionadas[21]. Para este, os meses são classificados como tratamentos e os anos como blocos. A ordenação é feita dentro de cada bloco, ao invés de ordenar todas as N observações. Pressupõe-se que os blocos são independentes entre si, ou seja, as observações de um ano independem das observações de outro ano.

A estatística T_5 de Friedman é dada por:

$$T_5 = \frac{12}{pk(k+1)} \sum_{j=1}^k R_{.j}^{*2} - 3p(k+1) \quad (2.16)$$

em que p é o número de blocos, k o número de tratamentos e $R_{.j}^*$ é o resultado da soma dos postos da j -ésima coluna, ou seja:

$$R_{.j}^* = \sum_{i=1}^p R_{ij}^* \quad (2.17)$$

onde R_{ij}^* é o posto de Y_{ij} dentro do bloco i , de 1 até k . A distribuição de T_5 pode ser aproximada por um χ^2 com $k - 1$ graus de liberdade.

Outro teste proposto para verificação de sazonalidade da série, mas que está no contexto da análise espectral, é o teste de Fisher, rotineiro a uma análise de variância[21]. O modelo

subjacente é dado por:

$$Y_{ij} = S_j + e_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, k, \quad (2.18)$$

e supondo $e_{ij} \sim N(0, \sigma^2)$, independentes. Sob a hipótese nula $H_0 : S_1 = \dots = S_k$, a estatística

$$T_5 = \frac{N - k}{k - 1} \frac{\sum_{j=1}^k n_j (\bar{Y}_{.j} - \bar{Y})^2}{\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2} \quad (2.19)$$

tem distribuição $F(k-1, N-k)$.

Vemos, pois, que condições para se aplicar este teste incluem a validade do modelo aditivo e normalidade dos resíduos[21].

2.1.3 MODELOS SAZONAIS

Como visto anteriormente, a sazonalidade faz parte da composição de uma série temporal e existem casos em que mesmo após o ajuste eliminando a componente sazonal determinística, ainda reste correlação significativa em:

- (i) “lags” de baixa ordem, indicando que os resíduos ainda são correlacionados, podendo-se ajustá-los através de um modelo ARIMA;
- (ii) “lag” múltiplos de período s , que significam a existência de uma sazonalidade estocástica, e a necessidade de se ajustar um modelo ARIMA sazonal, também conhecido como SARIMA.

A notação para essa classe de modelos é dada por $SARIMA(p,d,q)(P,D,Q)_s$ e definido por[21]:

$$\Phi(B^S)\phi(B)\Delta_S^D\Delta^d Z_t = \Theta(B^S)\theta(B)a_t \quad (2.20)$$

em que:

- $\Phi(B^S) = (1 - \Phi_1 B^S - \dots - \Phi_P B^{SP})$ é o operador regressivo sazonal de ordem P;
- $\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$ é o operador auto-regressivo de ordem p;
- $\Theta(B^S) = (1 - \Theta_1 B^S - \dots - \Theta_Q B^{SQ})$ é o operador de médias móveis de ordem Q;
- $\theta(B) = (1 - \theta_1 B - \dots - \theta_q B^q)$ é o operador de médias móveis sazonal de ordem q;
- $\Delta_S^D = (1 - B^S)^D$ é o operador diferença sazonal D, indicando o número de diferenças sazonais;
- $\Delta^d = (1 - B)^d$ é o operador diferença simples d, indicando o número de diferenças.

Usualmente, é adotado uma sazonalidade s igual a 12, equivalente a quantidade de meses em um ano.

2.1.4 CRITÉRIOS DE SELEÇÃO DE MODELOS

Após ajustados os modelos, é preciso definir um critério para seleção do modelo que melhor descreve os resultados observados. Existe na bibliografia diversos critérios de informação que levam em conta não apenas a qualidade do ajuste mas também penalizam a inclusão de parâmetros extras. A regra básica de seleção utilizando estes critérios consiste em selecionar o modelo cujo critério de informação calculado seja mínimo.

A comparação deve ser feita entre modelos ajustados no mesmo período amostral.

CRITÉRIO DE INFORMAÇÃO DE AKAIKE

O Critério de Informação de Akaike (AIC) é dado por:

$$AIC = -2\log(L) + 2m \quad (2.21)$$

em que L é a verossimilhança maximizada dos dados e m é o número de parâmetros do modelo. O critério de informação de Akaike (AIC) pode assumir qualquer valor, inclusive valor negativo, já que ele depende da forma da função de verossimilhança[2].

CRITÉRIO DE INFORMAÇÃO BAYESIANO

Já o Critério de Informação Bayesiano é definido por:

$$BIC = 2\log(L) + m\log(n) \quad (2.22)$$

em que L é a verossimilhança maximizada dos dados, m é o número de parâmetros do modelo e n é tamanho da amostra.

RAIZ QUADRÁTICA DO ERRO MÉDIO (RMSE)

Diferente dos critérios mencionados anteriormente, a Raiz Quadrática do Erro Médio (RMSE), definida por:

$$RMSE = \frac{\sum_{i=1}^N (O_i - P_i)^2}{N} \quad (2.23)$$

é comumente utilizada para expressar a acurácia dos resultados numéricos com a vantagem de que apresenta valores do erro nas mesmas dimensões da variável analisada. Pode ser interpretado como uma medida do desvio médio entre observado e predito.

ERRO PERCENTUAL MÉDIO ABSOLUTO (MAPE)

O Erro Percentual Médio absoluto de Previsão (sigla MAPE da expressão em inglês mean absolute percentual error) é dado por:

$$MAPE_{t,s} = \frac{\sum_{i=1}^8 (y_{ts+i} \hat{y}_{ts+i}) / y_{ts+i}}{s} \quad (2.24)$$

em que s é o horizonte de previsão, y_{ts+i} é o valor real das s últimas observações, $y_{ts+i}^{\hat{}}$ é o valor previsto em um modelo feito a partir dos dados sem as s últimas observações, e n é igual ao número de observações.

O MAPE é uma medida de erro comumente usada para aferir a qualidade do ajuste de um modelo, sendo considerado o melhor modelo, para fins de previsão, aquele que apresentar menor valor desta medida[21].

2.1.5 TESTES DE ADEQUAÇÃO DO MODELO

Feita a estimação dos modelos é necessário verificar se eles representam adequadamente os dados observados. Para isso, a literatura propõe alguns testes para validação, citados a seguir:

TESTE DE AUTOCORRELAÇÃO RESIDUAL

A partir do parâmetros estimados do modelo, é possível calcular os resíduos estimados por:

$$\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)W_t. \quad (2.25)$$

Para que um modelo seja adequado, \hat{a}_t devem estar próximos de a_t . Em outras palavras, \hat{a}_t e a_t devem ser aproximadamente não-correlacionados. Seja \hat{r}_k as autocorrelações dos resíduos \hat{a}_t , então $\hat{r}_k \sim N(0, 1/n)$, sempre sob a suposição que o modelo ajustado é apropriado. Uma boa indicação para verificar o comportamento de ruído branco em a_t é a comparação de \hat{r}_k com os limites $\pm \frac{2}{\sqrt{n}}$ [21].

TESTE DE BOX-PIERCE

Box e Pierce formularam um teste para as autocorrelações dos resíduos estimados, capaz de indicar se esses valores são muito altos[8]. A estatística é dada por:

$$Q(K) = n(n+2) \sum_{j=1}^k \frac{\hat{r}_j^2}{(n_j)}. \quad (2.26)$$

Um modelo apropriado terá aproximadamente uma distribuição qui-quadrado com $K - pq$ graus de liberdade. Grandes valores de $Q(K)$ rejeitam a hipótese nula de ruído branco para os resíduos.

TESTE DO PERIODOGRAMA ACUMULADO

O teste de periodograma acumulado constitui-se da comparação dos periodogramas acumulados da série dos resíduos estimados \hat{a}_t com a de um ruído branco a_t , a fim de verificar a existência de componentes periódicas na série dos resíduos estimados \hat{a}_t [13]. Denotando as observações de um processo estocástico por a_t , $t = 1, \dots, n$, um estimador do espectro $p_a(f)$ do

processo é dado por:

$$I_a(f_i) = \frac{1}{2\pi n} \left[\left(\sum_{t=1}^n a_t \cos\left(\frac{2\pi i}{n} t\right) \right)^2 + \left(\sum_{t=1}^n a_t \sin\left(\frac{2\pi i}{n} t\right) \right)^2 \right], 0 < (f_i) < 1/2 \quad (2.27)$$

Esse estimador do espectro é denominado periodograma. O espectro acumulado, $P_a(f)$, pode ser estimado por $\frac{1}{n} \sum_{i=1}^j I_q(f_i)$. Já a estimativa de $P_a(f)/\sigma_a^2$, denotada por C_{f_j} , é chamada de periodograma acumulado (normalizado). Prova-se que se a_t for ruído branco, então seu espectro é constante e igual a $2\sigma^2$ no intervalo $[0, 1/2]$. Analisando o gráfico de $C_{f_j} f_j$, se o mesmo aproxima de uma reta de inclinação 1 para $f_j \in [0, 1/2]$, então os resíduos podem ser considerados como ruído branco.

2.2 REDES NEURAIS ARTIFICIAIS (RNA)

Redes Neurais Artificiais são técnicas computacionais que utilizam modelos matemáticos para simular uma estrutura neural de seres vivos inteligentes e que adquirem conhecimento através de experiências. São constituídas por unidades de processamento, que por sua vez estão conectadas por canais de comunicação associados a pesos. A inteligência da estrutura se deve à iteração entre as unidades de entrada e seus respectivos pesos sinápticos. Essa técnica possui diversos modelos para regra de treinamento e é através deste modelo que a rede neural aprende e aprimora seu desempenho[15]. A figura abaixo representa a forma como é feita as iterações dentro de uma RNA.

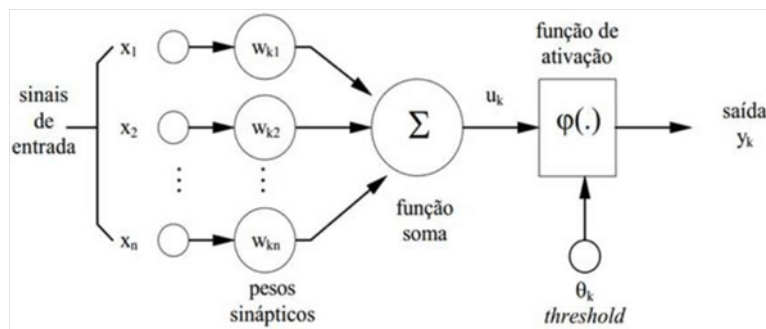


Figura 2.1: Estrutura de uma Rede Neural Artificial

Uma rede neural recorrente (RNN do inglês Recurrent Neural Network) é um tipo de rede neural artificial projetada para reconhecer padrões em sequências de dados, como texto, genomas, caligrafia, palavra falada ou dados de séries numéricas que emanam de sensores, bolsas de valores e agências governamentais[1].

A ideia principal dessas redes é que elas possam ser alimentadas reversamente com dados relevantes ao modelo. A mecânica consiste em cada neurônio de uma RNN ser conectado à neurônios das camadas anterior e seguinte. Em cada um é atribuído um peso para as entradas recebidas e deles subtraídos um valor conhecido como bias e submetidos a uma função de ativação. São esses pesos que representam o treinamento recebido e portanto constituem a memória da rede. Podemos entender que cada neurônio reconhece determinado tipo de situação

e para essa situação de entrada, esse neurônio está mais ativado, portanto para um problema complexo, são utilizadas grandes quantidades de neurônios e camadas[30].

Matematicamente, uma RNN pode ser definida como:

$$h^t = \sigma(W_h X + W_r y^{t-1}) \quad (2.28)$$

$$y = \sigma(W_y h^t) \quad (2.29)$$

em que h recebe o nome estado oculto, W são pesos, X entrada, y saída e σ a função sigmoide.

Segundo Haykin [15], o processo de realimentação proporciona grande influência na rede em termos de treinamento e desempenho. A Figura 2.2 apresenta um exemplo de rede neural artificial recorrente, destacando, principalmente, o processo de realimentação dos neurônios.

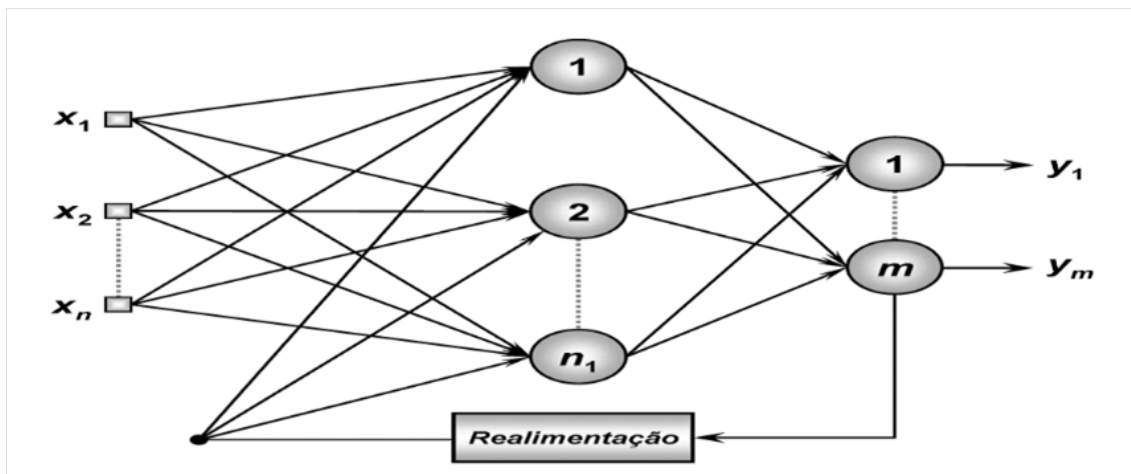


Figura 2.2: Estrutura de uma Rede Neural Recorrente [27]

2.2.1 REDES NEURAIIS DO TIPO LSTM

Unidades de Memória de Longo Prazo, ou simplesmente LSTM, são um tipo especial de RNN capaz de aprender dependências de longo prazo. Introduzido por Hochreiter Schmidhuber (1997)[16], e refinado e popularizado por muitas pessoas no trabalho seguinte[23].

O LSTM tem a capacidade de remover ou adicionar informações ao estado da célula, cuidadosamente regulado por estruturas chamadas de portões, que são uma forma opcional de permitir a passagem de informações. Eles são compostos de uma camada de rede neural sigmoide e uma operação de multiplicação pontual. Existem três portões, sendo eles[1][23]:

- **Portão de Esquecimento (Forget Gate):** As informações que não são mais úteis no estado da célula são removidas com o forget gate. Duas entradas: x_t (entrada no momento específico) e h_{t-1} (saída de célula anterior) são alimentadas ao gate e multiplicadas por matrizes de peso, seguidas pela adição do bias. O resultante é passado por uma função de ativação que fornece uma saída binária. Se para um determinado estado de célula a

saída for 0, a informação é esquecida e para a saída 1, a informação é retida para uso futuro. O processamento pode ser definido pela equação:

$$f_t = \sigma(W_f[h_{(t-1)}, x_{(t)}] + b_f) \quad (2.30)$$

- **Portão de Entrada (Input Gate):** A adição de informações úteis ao estado da célula é feita pelo input gate. Primeiro, a informação é regulada usando a função sigmoide que filtra os valores a serem lembrados de forma similar ao forget gate usando as entradas h_{t-1} e x_t . Então, um vetor é criado usando a função tanh que dá saída de -1 a $+1$, que contém todos os valores possíveis de h_{t-1} e x_t . Os valores do vetor e os valores regulados são multiplicados para obter as informações úteis. A equação é dada por:

$$i_t = \sigma(W_i[h_{(t-1)}, x_{(t)}] + b_i) \quad (2.31)$$

- **Portão de Saída (Output Gate):** A tarefa de extrair informações úteis do estado da célula atual para ser apresentadas como uma saída é feita pelo output gate. Primeiro, um vetor é gerado aplicando a função tanh na célula. Então, a informação é regulada usando a função sigmoide que filtra os valores a serem lembrados usando as entradas h_{t-1} e x_t . Os valores do vetor e os valores regulados são multiplicados para serem enviados como uma saída e entrada para a próxima célula.

$$o_t = \sigma(W_o[h_{(t-1)}, x_{(t)}] + b_o) \quad (2.32)$$

Conforme Dametto[12], os nós de memória de uma rede neural artificial do tipo LSTM são denominados de células, uma vez que esses tipos de estrutura possuem propriedades mais complexas que redes neurais recorrentes tradicionais pelo fato de aprenderem dependências no longo prazo, bem como selecionar informações que devem ser “esquecidas” a partir de um certo ponto. Esquemáticamente, a Figura 2.3 mostra uma rede neural artificial LSTM e seus itens constituintes.

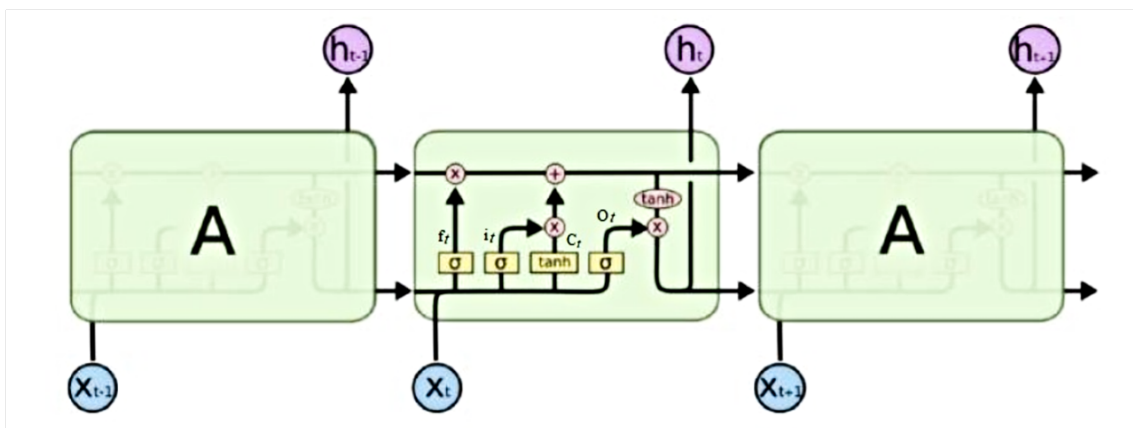


Figura 2.3: Estrutura de uma Rede LSTM

Em que:

- t : representa o tempo;
- C_t : é o estado da célula no instante de tempo t ;
- h_t : representa a saída da célula;
- x : é a amostra da sequência;
- f_t : é o portão de esquecimento;
- i_t : é o portão de entrada;
- O_t : é o portão de saída;

A simulação de uma rede LSTM depende de duas condições: ter dados estacionários, devendo eliminar tendência tomando diferenças sucessivas, e devem estar em uma escala de -1 a 1, a função de ativação das unidades da rede[9]. Além disso, deve estabelecer alguns parâmetros, que são descritos abaixo.

- Amostras de treino: o número de entradas utilizadas para treinar o modelo.
- Amostras de teste: o número de entradas utilizadas para testar o modelo.
- Épocas: o conjunto de treino é alimentado à rede tantas vezes quanto o número de épocas.
- Tamanho batch: o batch é o conjunto de entradas que a rede neural irá processar antes de ajustar os pesos do modelo. É muito impactante na velocidade do treinamento.
- Número de Neurônios: representa o número de unidades LSTM.
- Unidades anteriores: quantas unidades consecutivas compõem uma única entrada. Pode ser entendido como o tamanho da janela que se desloca à medida que os dados são lidos.
- Horizonte de predição: até quantas unidades após à entrada serão feitas previsões.
- Função de Perda: a principal função de qualquer rede neural é minimizar perdas, é por seus erros que ela aprende. Essa é a função utilizada para determinar quão ruim está a resposta da rede durante o treino, um exemplo seria o erro médio quadrático.
- Otimizador: um método numérico usado na otimização, para encontrar um mínimo da função (entrada, peso) \rightarrow (saída), normalmente usa-se um esquema iterativo onde em cada passo se toma a direção negativa do gradiente calculado durante a alimentação reversa da rede.
- Função de ativação: A função pela qual passa os valores de entrada de um neurônio submetidos ao peso, cujo resultado representa o estado de ativação do neurônio.

Por fim, para avaliar a precisão do modelo e facilitar comparações com outros tipos de análises, é bastante comum o uso de parâmetros de erro como Raiz Quadrática do Erro Médio (RMSE), por exemplo, o qual será utilizado posteriormente para avaliar os resultados obtidos.

3. METODOLOGIA

A base de dados utilizada neste trabalho se refere à série histórica mensal correspondente ao Índice de Preços ao Consumidor Amplo (IPCA), obtidos no site do Instituto Brasileiro de Geografia e Estatística (IBGE), através do Sistema IBGE de Recuperação Automática (SIDRA)[26].

Considerou-se dois cenários para as análises e previsões, levando em conta fatores externos que podem influenciar diretamente o valor do IPCA:

- **Cenário A:** Período de Dezembro de 1999 a Abril de 2020 para ajustar os modelos e as previsões foram feitas para o período de Maio de 2020 a Abril de 2021 (total de 12 meses de previsão). Neste cenário está contemplado o período em que teve início a pandemia da COVID-19, que gerou impactos relevantes na economia do país.
- **Cenário B:** Período de Dezembro de 1999 a Dezembro de 2018 para ajustar os modelos e as previsões foram feitas para o período de Janeiro de 2019 à Dezembro de 2019 (total de 12 meses de previsão). Para este cenário, desconsiderou-se o período da pandemia da COVID-19, de modo que os dados tivessem menor interferência de fatores externos e permitissem estimar um modelo que se adeque melhor aos valores observados.

Para construção dos modelos, foi aplicada a técnica de Box & Jenkins para estimação da modelagem SARIMA e a técnica de Redes Neurais Recorrentes, chamada LSTM, para cada cenário citado anteriormente.

Alguns gráficos das séries foram projetados com o auxílio da ferramenta LibreOffice Calc [14].

MODELAGEM SARIMA

Com o auxílio do software R[29], foram realizados os cálculos e a construção dos modelos SARIMA. Optou-se pela utilização de um software livre e com recursos computacionais que apresentam funcionalidades específicas para séries temporais univariadas, especialmente o pacote *forecast*[17], que auxilia principalmente nas etapas de identificação e estimação dos melhores modelos.

Para uma análise inicial sobre o comportamento da série, foi construído o gráfico com o valor do IPCA acumulado mensal, para proporcionar uma inspeção visual.

A análise gráfica da função de autocorrelação foi realizada para averiguar a estacionariedade da série. Em seguida foi realizado um teste para verificar se os dados apresentam aleatoriedade ou não, e em caso negativo, aplicar a transformação.

Para identificar presença de tendência na série, foi utilizado o Coeficiente de Correlação de Spearman, o teste de Wald-Wolfowitz (teste de sequências) e o teste de Cox Stuart (teste do sinal), este último por meio do comando `cox.stuart.test()` pertencente ao pacote `randtests`. Os testes identificaram a presença da tendência, e em função disso, a tendência foi eliminada tomando uma diferença da série, aplicando o comando `diff()`.

Após a tendência ser eliminada aplicou-se o teste de Friedman e o teste de Fisher para averiguar a presença da sazonalidade, utilizando o comando `friedman.test()` do pacote `stats` e o comando `fisher.g.test()` do pacote `GeneCycle`, respectivamente, em que ambos confirmaram a sua existência.

Através do comando `auto.arima()` do pacote `forecast`, foram definidos possíveis modelos para a série temporal. Ao todo, foram testados 15 modelos SARIMA, representados por SARIMA(p, d, q)(P,D,Q). Utilizou-se de técnicas de comparação de modelos para escolher o melhor modelo possível, por meio dos indicadores: Critério de Informação de Akaike (AIC), Critério de Informação Bayesiano (BIC), Raiz Quadrática do Erro Médio (RMSE) e Erro Percentual Médio absoluto (MAPE). Também foi levado como critério de avaliação a existência de parâmetros não significativos, o teste de autocorrelação residual, o teste de Box-Pierce e o teste do periodograma acumulado.

Após a definição dos melhores modelos, fez-se a previsão para $t = 12$ meses, e plotou-se o gráfico da série com os dados estimados e observados. Todos os procedimentos citados foram aplicados nos dois cenários propostos.

LSTM

As simulações de LSTM foram feitas utilizando a linguagem Python[25], através do Google Colaboratory[10], que é um ambiente de notebooks Jupyter que não requer configuração e é executado na nuvem. Os pacotes utilizados para as análises foram: pacote Keras, uma biblioteca redes neurais, além das bibliotecas Pandas para estruturação dos dados, Matplotlib para a construção de gráficos, scikit-learn para regressão linear e normalização dos dados e NumPy pela estrutura de vetor e funções matemáticas. A escolha dessa ferramenta se deu por conta de ser um software livre e gratuito, pela grande quantidade de publicações que fazem uso da ferramenta, além de ter uma comunidade ativa que consegue apoiar na solução de problemas. O algoritmo em que as análises foram baseadas foi obtido no site Machine Learning Mastery[9].

Inicialmente, tomou-se uma diferença nos dados para eliminar tendência e torná-los estacionários. Logo após, realizou-se uma transformação escalar nos dados de forma a ficarem entre os valores -1 e 1, através do comando `MinMaxScaler` da biblioteca `sklearn`.

Em seguida, ajustou-se um modelo de rede LSTM aos dados de treinamento. Para essa aplicação optou-se por definir um lote de tamanho 1, também chamado de aprendizado online, pois os pesos da rede serão atualizados durante o treinamento após cada padrão de treinamento. A rede utilizou uma função de perda de erro quadrático médio e o algoritmo de otimização ADAM, que melhora funções estocásticas.

A partir dos modelos ajustados, realizou-se as previsões para os últimos 12 meses. Para as

análises, adotou-se duas técnicas diferentes para previsão, visto que o objetivo na sequência é comparar o desempenho das RNN frente aos modelos SARIMA. Definimos da seguinte forma:

- **LSTM Sem Reposição:** Neste tipo de modelo, a previsão dos 12 meses será feita de uma só vez a partir de uma entrada t . Para tal, fixamos o número de sequências de saída em 12 (n_{seq}) e o $n_{test} = 1$, sendo esse ponto t o último mês anterior ao período separado para previsão (No cenário A, de Maio/2020 à Abril/2021, e no cenário B de Janeiro/2019 à Dezembro/2019). Desta forma, as saídas se aproximam do formato que temos em modelos SARIMA, tornando mais justa a comparação entre os modelos.
- **LSTM Com Reposição:** Neste tipo de modelo, a previsão terá recorrência mensal, com 1 saída para cada entrada num período de 12 meses. Para tal, fixamos o número de sequências de saída ($n_{seq} = 1$) e o $n_{test} = 12$, sendo esse ponto t o último mês anterior ao período separado para previsão (No cenário A, de Maio/2020 à Abril/2021, e no cenário B de Janeiro/2019 à Dezembro/2019). Neste formato, por ter a reposição dos dados, não se assemelha à forma como estabelecemos nos modelos SARIMA.

Na sequência das previsões, inverteu-se as transformações de escala e diferença realizadas nos dados anteriormente e obteve-se o valor da Raiz Quadrática do Erro Médio (RMSE), para avaliar a acurácia das previsões.

É preciso definir a quantidade de neurônios e a quantidade de épocas a serem adotadas no modelo, porém não está claro na literatura uma regra clara de como estimar essas quantidades. Logo, para o presente trabalho, optou-se por realizar o ajuste dos modelos para as seguintes quantidades de neurônios e épocas:

- **Neurônios:** 7, 25, 50, 100 e 200.
- **Épocas:** 100, 300, 500 e 1000 épocas.

que resultou em 40 combinações de modelos, que foram estimados para cada técnica de previsão (LSTM Com Reposição e LSTM Sem Reposição), obtendo ao final 80 modelos distintos. O critério de seleção do melhor modelo foi a partir do que apresentou o menor valor para o RMSE para cada técnica de previsão.

COMPARAÇÃO ENTRE TÉCNICAS

Para comparação entre os melhores modelos propostos por cada técnica (Modelagem SARIMA, LSTM Sem Reposição e LSTM Com Reposição), fez-se uso do cálculo da Raiz Quadrática do Erro Médio (RMSE), de tal forma a escolher aquele que apresentar o menor valor.

4. RESULTADOS

4.1 CENÁRIO A

O gráfico da série histórica mensal do IPCA no período de Dezembro/1999 à Abril/2021 é apresentado na Figura 4.1. Não fica evidente pela análise gráfica se os dados apresentam sazonalidade. Também pode-se observar que a série é não estacionária, uma vez que apresenta uma tendência negativa, indicando uma queda no valor da inflação nos últimos meses. Além disso, é possível verificar uma alta brusca no nível da série, aproximadamente entre os anos de 2002 e 2003, e no início de 2021. Este último em decorrência da pandemia da COVID-19.

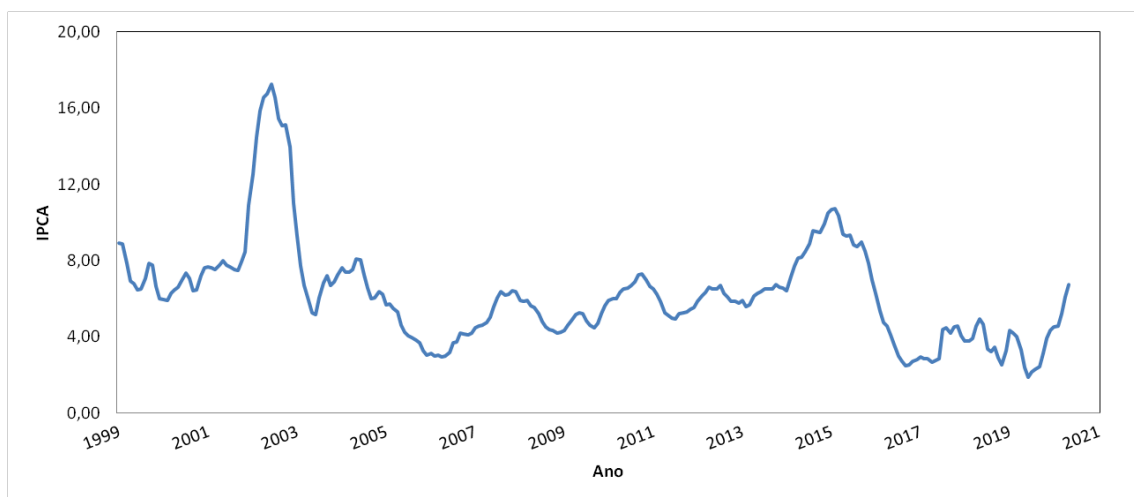


Figura 4.1: IPCA acumulado mensal de Dezembro/1999 à Abril/2021

4.1.1 MODELOS SARIMA

Inicialmente para as análises de modelos SARIMA, propostos por Box & Jenkins[8], obteve-se o gráfico da função de autocorrelação (FAC) ilustrado na Figura 4.2 para identificar graficamente se a série é estacionária. Em caso positivo, o gráfico apresentará um comportamento de decaimento para zero relativamente rápido. Observa-se que não é o caso da série estudada, indicando que os dados não são estacionários.

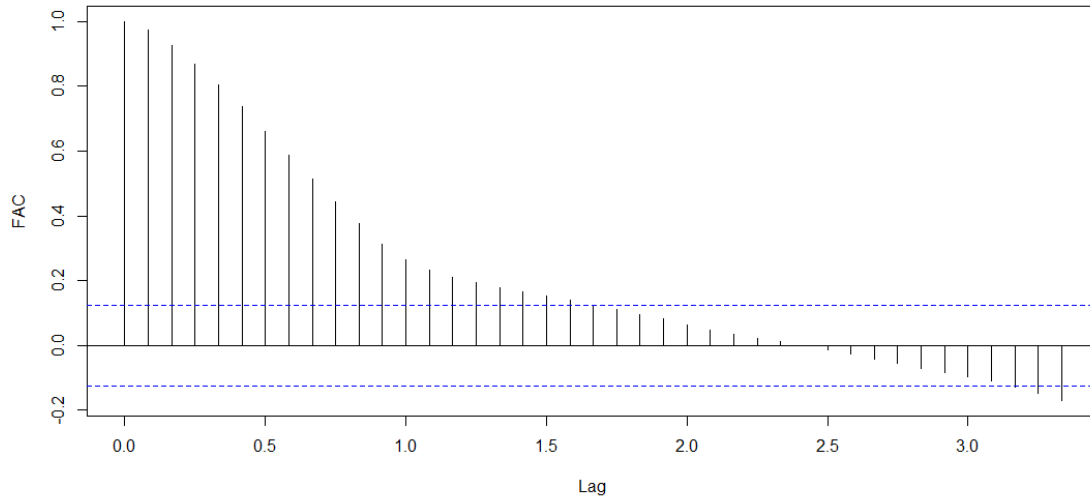


Figura 4.2: Função de Autocorrelação da série (Cenário A)

A partir da Figura 4.3, foi plotada a média em função da amplitude. Essas medidas foram calculadas dividindo a série original em vários grupos composto de duas observações. Sendo assim, foi possível identificar que um modelo aditivo é adequado a ser adotado para estimação, uma vez que os componentes da série atuam de modo absoluto e independente entre si. Além disso, verificou-se que não é necessário o uso de transformações, já que a amplitude não depende da média.

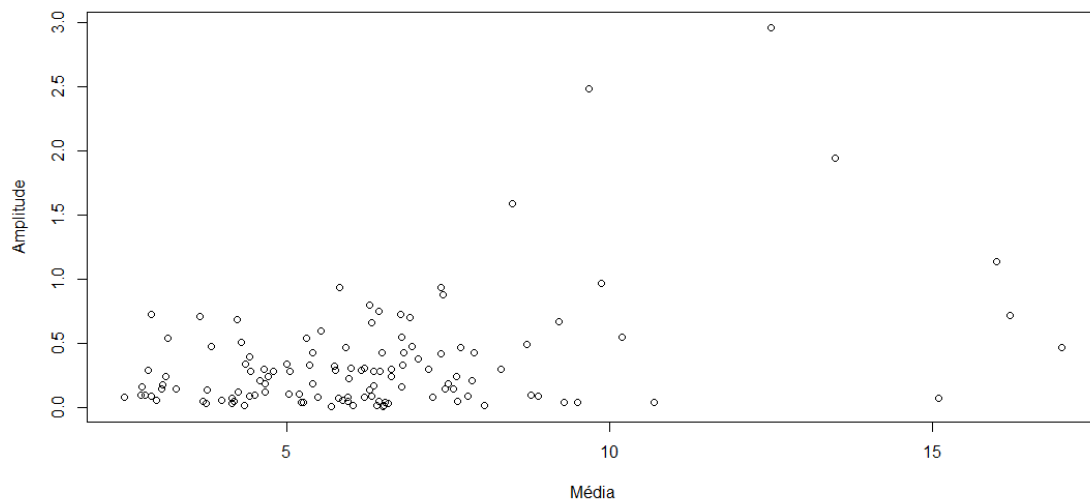


Figura 4.3: Distribuição Média pela Amplitude de grupos de observações da série (Cenário A)

Na Tabela 4.1 tem-se o resultado dos testes para tendência. Como hipótese nula dos testes considera-se a não existência de tendência. Desta forma, verificou-se que a hipótese nula foi rejeitada ao nível de significância de 5% para os três testes realizados, sendo necessário realizar uma diferença na série para eliminar a tendência.

Tabela 4.1: Teste para verificação de tendência na série (Cenário A)

Teste	P-valor	Conclusão
Wald-Wolfowitz	$<2,2e-16$	Rejeita-se H_0
Cox-Stuart	0,002662	Rejeita-se H_0
Correlação de Spearman	$1,799e-11$	Rejeita-se H_0

Identificada a tendência, tomou-se uma diferença na série a fim de eliminar esse comportamento e tornar os dados estacionários. Aplicando novamente os testes de tendência, constatou-se a eliminação dessa componente. O gráfico da série com a primeira diferença pode ser visualizado na Figura 4.4, sendo possível observar média constante com variabilidade baixa, levando a considerar que não há a necessidade de realizar uma segunda diferença. Ainda na validação da tendência da série diferenciada, é possível examinar pela Figura 4.5 que as defasagens significativas, na FAC, seguem um padrão, reforçando a escolha pelo ajuste de um modelo do tipo autorregressivo integrado de médias móveis sazonal e o primeiro “lag” significativo na FACp dá sinais da presença do parâmetro regressivo p no modelo.

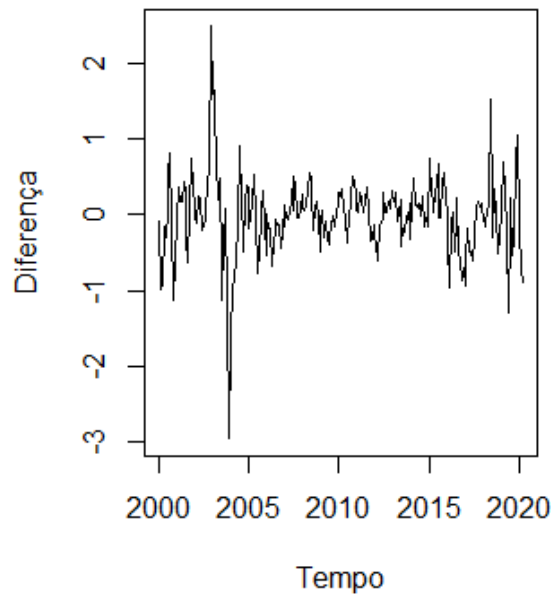


Figura 4.4: Série original com uma diferença (Cenário A)

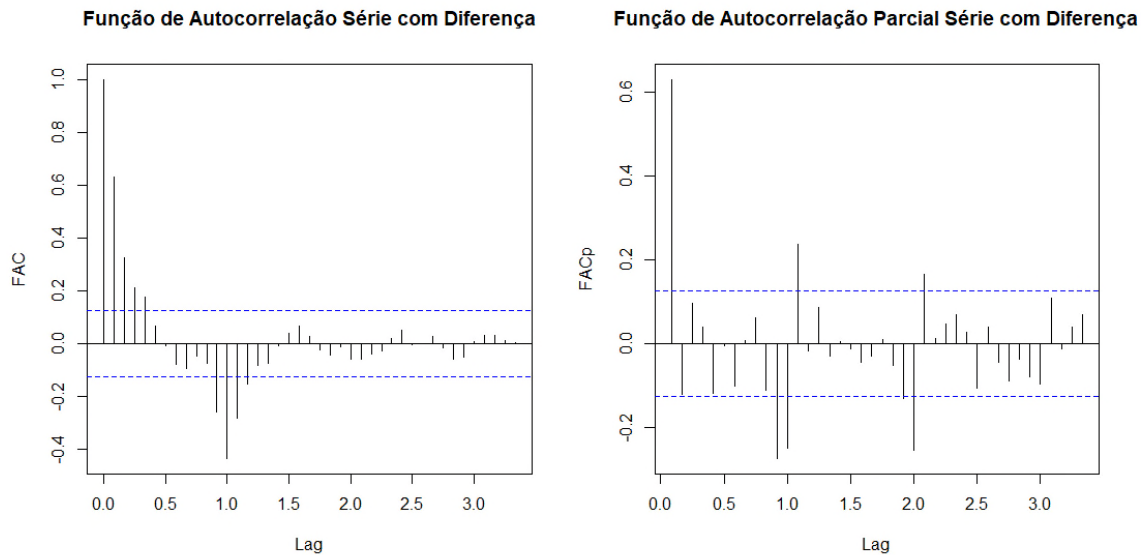


Figura 4.5: FAC e FACp da série diferenciada (Cenário A)

Eliminada a tendência na série, aplicou-se os testes de sazonalidade de Friedman e Fisher na série diferenciada. Como é possível observar na Tabela 4.2, ambos os testes indicam a existência da componente de sazonalidade, reforçando a necessidade de ajuste de modelos do tipo sazonais.

Tabela 4.2: Teste para verificação de sazonalidade na série (Cenário A)

Teste	P-valor	Conclusão
Friedman	2,2e-16	Rejeita-se H0
Fisher	2,228892e-21	Rejeita-se H0

Posterior à detecção dos componentes de tendência e sazonalidade da série, iniciou-se a etapa de estimação do modelo. Para ajuste dos modelos considerou-se uma diferença de ordem 1 na série para retirar a tendência. Com auxílio do comando *auto.arima()*, foram propostos 15 modelos, representados na Tabela 4.3.

Tabela 4.3: Ajustes de modelos SARIMA propostos

Modelo	Parâmetros Significativos
Modelo 1 - SARIMA(1,1,0)(2,0,1) _[12]	Não
Modelo 2 - SARIMA(1,1,0)(1,0,2) _[12]	Não
Modelo 3 - SARIMA(1,1,1)(2,0,1) _[12]	Não
Modelo 4 - SARIMA(1,1,0)(2,0,2) _[12]	Não
Modelo 5 - SARIMA(2,1,1)(2,0,1) _[12]	Não
Modelo 6 - SARIMA(1,1,0)(1,0,1) _[12]	Não
Modelo 7 - SARIMA(1,1,0)(2,0,0) _[12]	Sim
Modelo 8 - SARIMA(1,1,0)(1,0,0) _[12]	Sim

Modelo 9 - SARIMA(0,1,1)(2,0,1)_[12]	Não
Modelo 10 - SARIMA(0,1,0)(2,0,1)_[12]	Não
Modelo 11 - SARIMA(1,1,0)(3,0,0)_[12]	Sim
Modelo 12 - SARIMA(3,1,0)(3,0,0)_[12]	Não
Modelo 13 - SARIMA(1,1,0)(4,0,0)_[12]	Não
Modelo 14 - SARIMA(2,1,0)(1,0,0)_[12]	Não
Modelo 15 - SARIMA(1,1,1)(1,0,0)_[12]	Não

Identificados os parâmetros dos modelos propostos, adotou-se como critério inicial de seleção a significância de todos os parâmetros. Para tanto, selecionou-se apenas os modelos 7 (SARIMA(1,1,0)(2,0,0)_[12]), 8 (SARIMA(1,1,0)(1,0,0)_[12]) e 11 (SARIMA(1,1,0)(3,0,0)_[12]). Para estes, obteve-se os valores para os critérios de seleção: RMSE, MAPE, AIC e BIC, retratados na Tabela 4.4.

Tabela 4.4: Critérios de seleção dos modelos SARIMA propostos (Cenário A)

Modelo	RMSE	MAPE	AIC	BIC
Modelo 7 - SARIMA (1,1,0)(2,0,0) _[12]	2,6727	44,4835	177,4923	191,4810
Modelo 8 - SARIMA(1,1,0)(1,0,0) _[12]	2,5194	41,2266	203,4390	213,9305
Modelo 11 - SARIMA(1,1,0)(3,0,0) _[12]	2,6112	42,5610	167,8535	185,3393

Para o Modelo 8 (SARIMA(1,1,0)(1,0,0)_[12]) obteve-se os menores valores para o RMSE e o MAPE, enquanto que para o Modelo 11 (SARIMA(1,1,0)(3,0,0)_[12]) os menores valores para o AIC e BIC. Pelo Princípio da Parcimônia e por apresentar menores erros de previsão, optou-se por adotar o Modelo 8, uma vez que este apresenta também uma quantidade menor de parâmetros. Posto isto, os parâmetros estimados são apresentados na Tabela 4.5.

Tabela 4.5: Parâmetros estimados do modelo SARIMA (1,1,0)(1,0,0)_[12]

Modelo	Parâmetro	Estimativa	Desvio-Padrão
SARIMA(1,1,0)(1,0,0) _[12]	ϕ_1	0,6452	0,0490
	Φ_1	-0,5142	0,0595

Para validação da adequabilidade do modelo, foram avaliadas as funções de autocorrelação (FAC) e autocorrelação parcial (FACp) dos resíduos, ilustradas na Figura 4.6. É possível observar a existência de até 2 “lags” significativos, indicando que os resíduos são não correlacionados. Para o teste de Box-Pierce, que permite avaliar a hipótese nula de ruído branco para os resíduos, obteve-se o p-valor de 0,5145, concluindo que os resíduos são um processo de ruído branco, mesma conclusão obtida pelo teste do periodograma acumulado, apresentado na Figura 4.7.

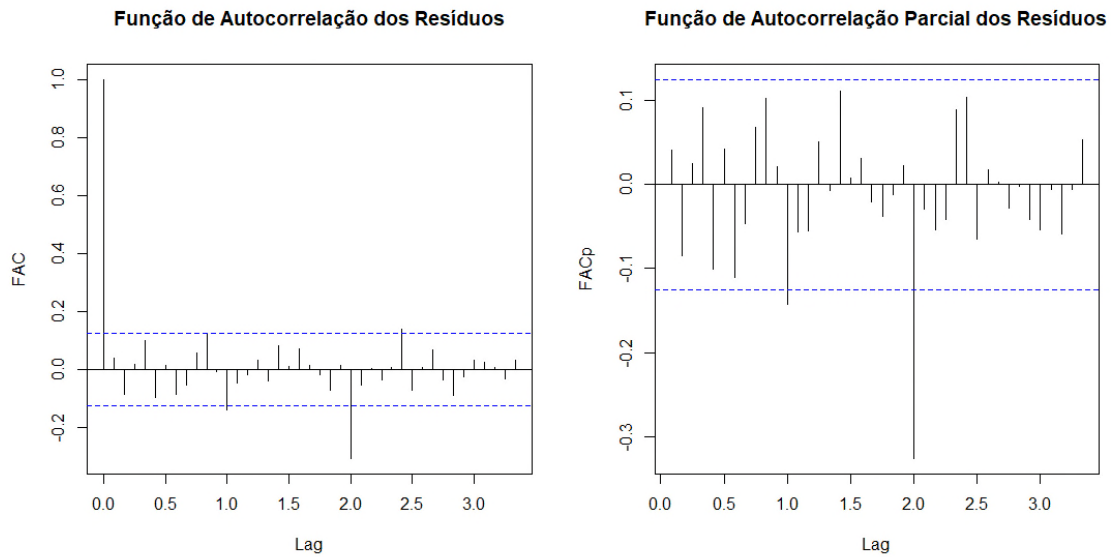


Figura 4.6: Gráficos de FAC e FACp dos Resíduos (Cenário A)

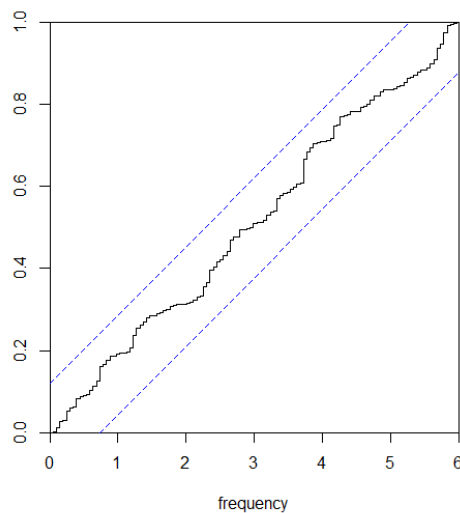


Figura 4.7: Periodograma Acumulado (Cenário A)

Na sequência, validado o modelo proposto, obteve-se as previsões para o período de Maio/2020 à Abril/2021. Os valores obtidos foram representados na Figura 4.8, comparando-se com os valores observados para o mesmo período. Observa-se que os valores previstos estão bem abaixo dos reais. Isto pode ser justificado por conta da pandemia da COVID-19, que teve seu início em 11 de março de 2020, de acordo com a Organização Mundial de Saúde[24]. Por se tratar de uma série econômica, fatores como esse influenciam na acertabilidade do modelo, visto que não se torna fácil a previsão destas situações.

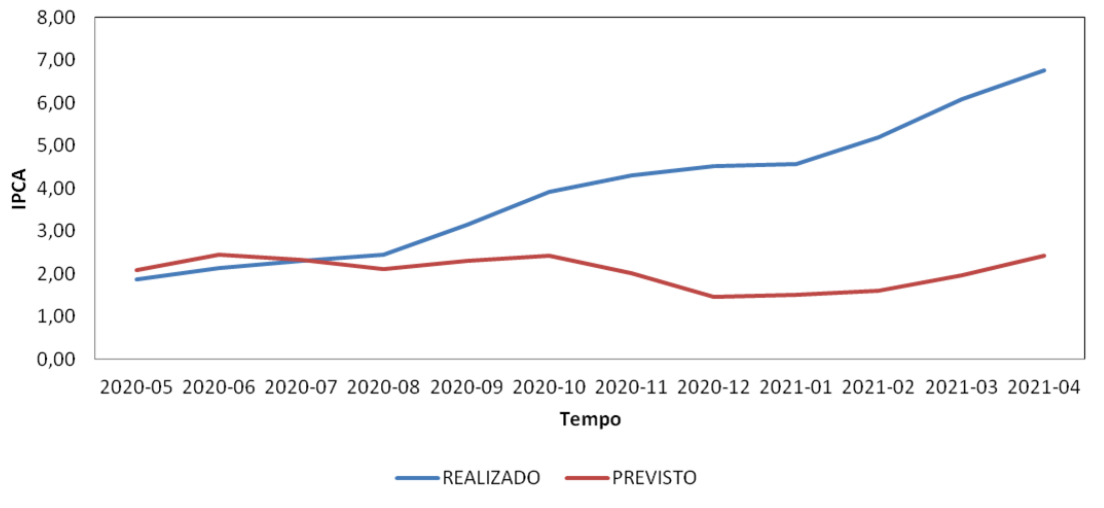


Figura 4.8: Previsões do modelo SARIMA (Cenário A)

Os valores previstos bem como os respectivos erros de previsão foram apresentadas na Tabela 4.6.

Tabela 4.6: Previsão realizada utilizando modelo SARIMA e erro de previsão (Cenário A).

Período	Valor Real	Valor Previsto	Erro de Previsão
2020-05	1,88	2,08274	0,20274
2020-06	2,13	2,44851	0,31851
2020-07	2,31	2,33363	0,02363
2020-08	2,44	2,10175	0,33825
2020-09	3,14	2,29951	0,84050
2020-10	3,92	2,42791	1,49209
2020-11	4,31	2,01925	2,29075
2020-12	4,52	1,46297	3,05703
2021-01	4,56	1,51083	3,04917
2021-02	5,20	1,59445	3,60555
2021-03	6,10	1,95379	4,14621
2021-04	6,76	2,41288	4,34712

Bonno (2014)[7] em seu estudo propôs um ajuste $SARIMA(0,1,1)(1,0,1)_{[12]}$ como método de estimação do valor da inflação, no período compreendido entre março de 2003 à março de 2013, reservando o período de abril de 2012 a março de 2013 para validação do modelo. Além disso, comparou os resultados a um modelo estrutural que avalia outras variáveis em função da inflação. Ela concluiu que os resultados obtidos pelo método proposto por Box & Jenkins se mostraram superiores aos resultados do modelo estrutural, semelhante ao comportamento averiguado no presente estudo.

4.1.2 MODELOS LSTM

Assim como proposto na sessão anterior, foram ajustados 40 modelos LSTM diferentes, sendo 20 para modelos do tipo *LSTM Sem Reposição* e 20 para modelos do tipo *LSTM Com*

Reposição, na intenção de definir a melhor combinação de neurônios e épocas. Os demais parâmetros e manipulações são dados por:

1. Tomou-se uma diferença nos dados para eliminar tendência e torná-los estacionários;
2. Realizou-se uma transformação escalar nos dados de forma a ficarem entre os valores -1 e 1, através do comando *MinMaxScaler* da biblioteca *sklearn*;
3. Ajustou-se um modelo de rede LSTM aos dados de treinamento com lote de tamanho 1, função de perda de erro quadrático médio e o algoritmo de otimização ADAM;
4. Realizou-se as previsões para o período de Maio/2020 à Abril/2021;
5. Inverteu-se as transformações de escala e diferença realizadas nos dados anteriormente;
6. Obteve-se o valor da Raiz Quadrática do Erro Médio (RMSE)

Os resultados obtidos para cada tipo de aplicação são dados a seguir.

LSTM SEM REPOSIÇÃO

Para este tipo de aplicação, foram testadas as combinações de quantidade de neurônios (7, 25, 50, 100 e 200) e épocas (100, 300, 500 e 1000). Para todos os modelos ajustados, obteve-se o valor do RMSE, a fim de selecionar o modelo com o melhor ajuste. A Tabela 4.7 apresenta todos os valores para cada combinação. O modelo selecionado foi o LSTM Sem Reposição com 25 Neurônios e 300 Épocas.

Tabela 4.7: Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Sem Reposição (Cenário A)

Neurônios	Épocas			
	100 Épocas	300 Épocas	500 Épocas	1000 Épocas
7	2,765789	3,472479	2,549114	3,223327
25	2,806572	2,548890	2,752101	4,079657
50	3,315780	2,956620	2,991710	3,851407
100	3,463543	2,911312	3,179212	3,651831
200	3,390992	3,216188	3,630765	4,576056

Definido o modelo mais adequado, projetou-se os valores previstos em um gráfico comparando com os valores reais, que pode ser observado na Figura 4.9. Ainda sobre a previsão desse modelo, é possível visualizar na Tabela 4.8 os valores previstos para cada mês, no período de Maio/2020 à Abril/2021, comparados aos valores reais e, os respectivos, erros de previsão.

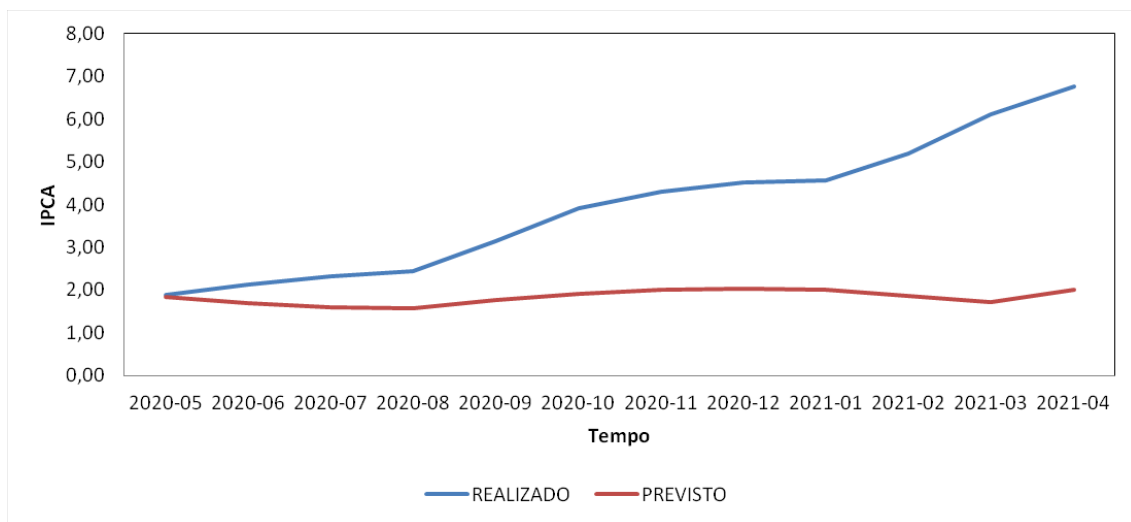


Figura 4.9: Previsões do modelo LSTM Sem Reposição (Cenário A)

Tabela 4.8: Previsão realizada utilizando modelo LSTM Sem Reposição (Cenário A)

Período	Valor Real	Valor Previsto	Erro de Previsão
2020-05	1,88	1,84187	0,03813
2020-06	2,13	1,68897	0,44103
2020-07	2,31	1,60720	0,70280
2020-08	2,44	1,58129	0,85871
2020-09	3,14	1,76507	1,37494
2020-10	3,92	1,91536	2,00464
2020-11	4,31	2,00215	2,30785
2020-12	4,52	2,03925	2,48075
2021-01	4,56	2,01254	2,54746
2021-02	5,20	1,87354	3,32646
2021-03	6,10	1,72928	4,37073
2021-04	6,76	2,01792	4,74208

Como é possível observar, as previsões para o período de Maio/2020 à Abril/2021 não foram muito próximas aos reais, e o erro de previsão tem aumento constante conforme vai passando os meses, da mesma forma que ocorreu nas previsões do modelo SARIMA $(1,1,0)(1,0,0)$ _[12].

LSTM COM REPOSIÇÃO

Para esta técnica, ajustou-se os 20 modelos LSTM Com Reposição para as combinações de Neurônios e Épocas listadas anteriormente. Calculou-se o RMSE a fim de definir o melhor ajuste de modelo para este cenário, representados na Tabela 4.9.

Tabela 4.9: Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Com Reposição (Cenário A)

Neurônios	Épocas			
	100 Épocas	300 Épocas	500 Épocas	1000 Épocas
7	0,361231	0,404639	0,506719	0,637176
25	0,385922	0,453928	0,636163	0,867946
50	0,429730	0,766542	0,573208	0,856281
100	0,492515	0,797309	1,334063	1,386368
200	0,379334	0,433664	0,538299	0,516772

Com base nos valores obtidos da raiz do erro quadrático médio, o modelo mais adequado é o LSTM Com Reposição com 7 Neurônios e 100 Épocas. Definido o melhor ajuste, obteve-se um gráfico comparativo dos valores reais e estimados, representado na Figura 4.10. Por fim, calculou-se o erro de previsão mensal do período de Maio/2020 à Abril/2021 na Tabela 4.10.

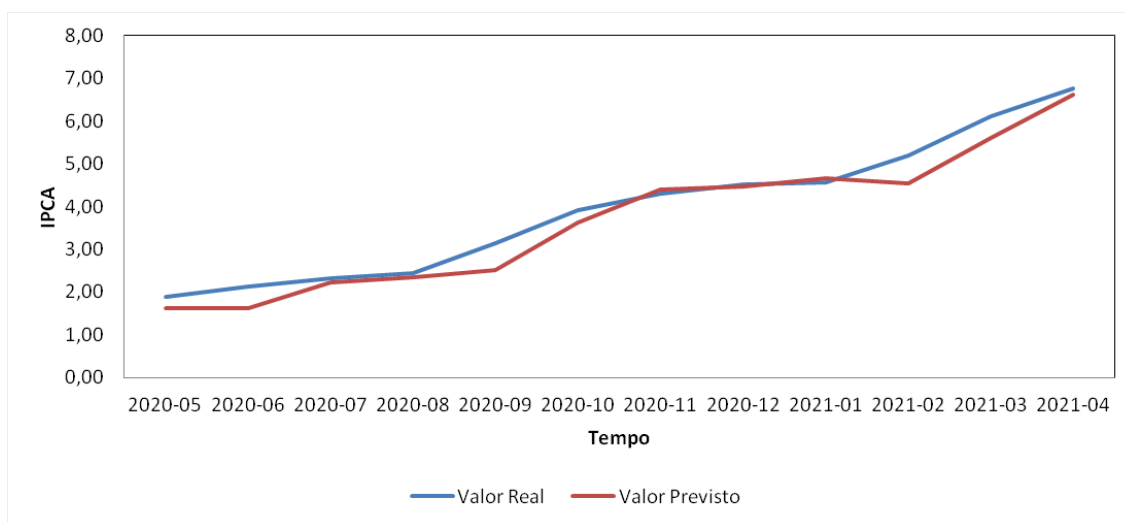


Figura 4.10: Previsões do modelo LSTM Com Reposição (Cenário A)

Tabela 4.10: Previsão realizada utilizando modelo LSTM Com Reposição e erro de previsão (Cenário A)

Período	Valor Real	Valor Previsto	Erro de Previsão
2020-05	1,88	1,62810	0,25190
2020-06	2,13	1,62152	0,50848
2020-07	2,31	2,23554	0,07446
2020-08	2,44	2,33935	0,10065
2020-09	3,14	2,50340	0,63660
2020-10	3,92	3,61530	0,30470
2020-11	4,31	4,39424	0,08424
2020-12	4,52	4,47469	0,04531
2021-01	4,56	4,65160	0,09160
2021-02	5,20	4,53878	0,66122
2021-03	6,10	5,59565	0,50435
2021-04	6,76	6,61523	0,14477

Observa-se pela Tabela 4.10 que para esse ajuste, os erros de previsão foram baixos, ou seja, os valores estimados se aproximam bastante dos valores reais, comprovando a eficácia de modelos de RNN para estimação de valores a curto prazo.

4.1.3 COMPARAÇÃO ENTRE OS MODELOS SARIMA E LSTM

Por fim, ao ajustar os modelos por meio dessas técnicas, é preciso estabelecer um critério que avalie qual modelo apresentou melhor desempenho. Para isso, utilizou-se o RMSE de todos os modelos obtidos para eleger o ajuste mais adequado para cada tipo de situação. Na Tabela 4.12 é possível observar o RMSE para cada um dos melhores modelos selecionados anteriormente.

Tabela 4.12: Valores de RMSE dos modelos selecionados de cada técnica estudada (Cenário A)

Modelo	RMSE
Modelo 8 - SARIMA(1,1,0)(1,0,0)	2,5194
LSTM Sem Reposição - 25 Neurônios e 300 Épocas	2,5489
LSTM Com Reposição - 7 Neurônios e 100 Épocas	0,3612

Obteve-se também a projeção gráfica das estimativas de todos os modelos selecionados para cada técnica comparados ao valor real da inflação no período de Maio/2020 à Abril/2021, representado na Figura 4.11.

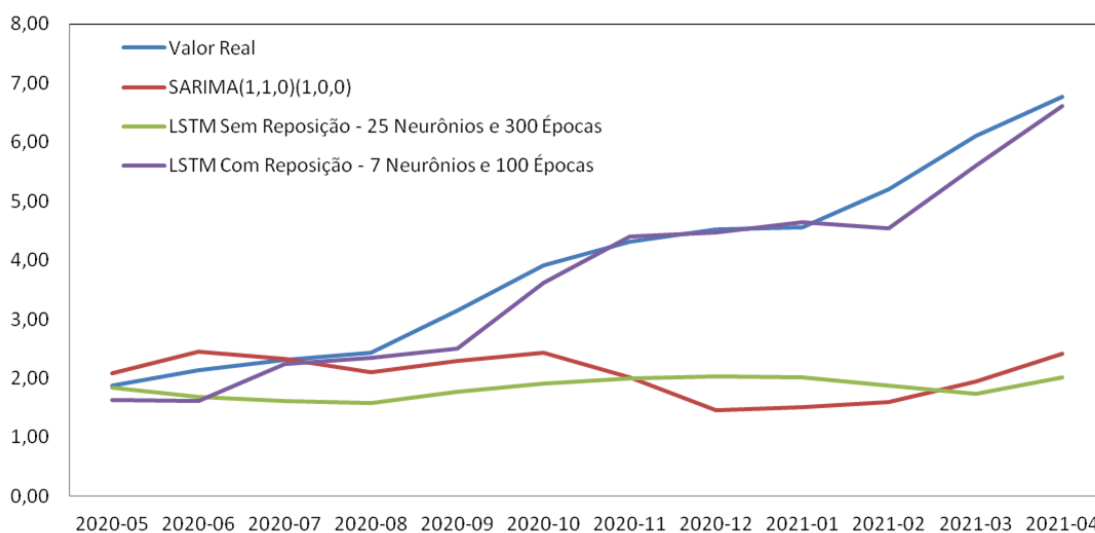


Figura 4.11: Série Original e Previsões dos modelos (Cenário A)

Comparando o RMSE tem-se que o modelo que mais se aproximou dos valores observados é o LSTM Com Reposição - 7 Neurônios e 100 Épocas, porém a técnica deste não é a mesma da técnica dos demais modelos (SARIMA(1,1,0)(1,0,0) e LSTM Sem Reposição - 25 Neurônios e 300 Épocas), os quais buscam definir um modelo que seja capaz de estimar os próximos 12 meses de uma só vez, sem que haja a realimentação de informação. Comparando o modelo SARIMA(1,1,0)(1,0,0) com o LSTM Sem Reposição - 25 Neurônios e 300 Épocas, temos que a

técnica proposta por Box & Jenkins[8] obteve melhores resultados, mesmo que ainda distantes do real.

4.2 CENÁRIO B

Tendo em vista o cenário político e econômico vivido no Brasil e no mundo no período da pandemia da COVID-19, os valores da inflação tendem a ser afetados. Por conta disto, buscou-se a estimação de modelos sem a interferência de fatores que não são possíveis de se prever. As análises realizadas neste cenário contemplam o período de Dezembro/1999 à Dezembro/2018, e o período de Janeiro/2019 à Dezembro/2019 separado para validação dos modelos. A Figura 4.12 ilustra o comportamento dos dados neste intervalo.

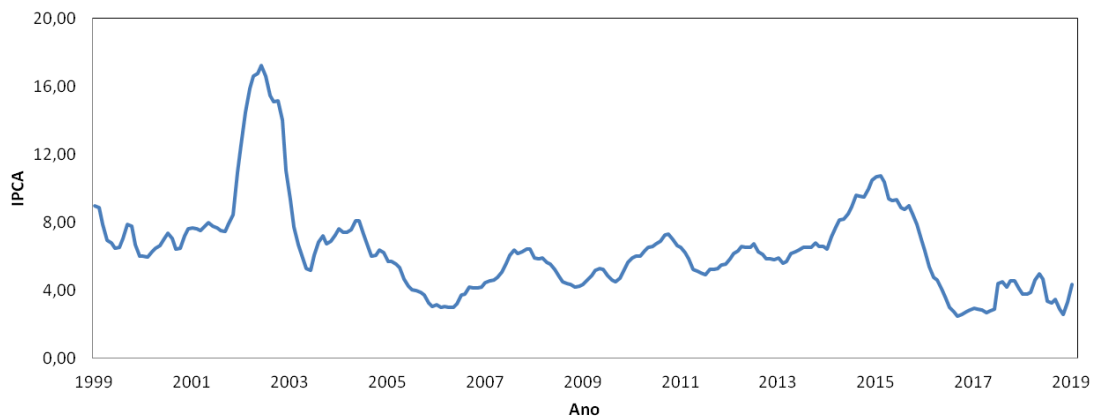


Figura 4.12: IPCA acumulado mensal de Dezembro/1999 à Dezembro/2019

4.2.1 MODELOS SARIMA

Primeiramente, obteve-se o gráfico da função de autocorrelação (FAC) ilustrado na Figura 4.13 para avaliar graficamente se a série estudada apresenta estacionariedade. Constata-se que a série não possui comportamento de estacionariedade, uma vez que não apresenta um decaimento para zero de forma brusca.

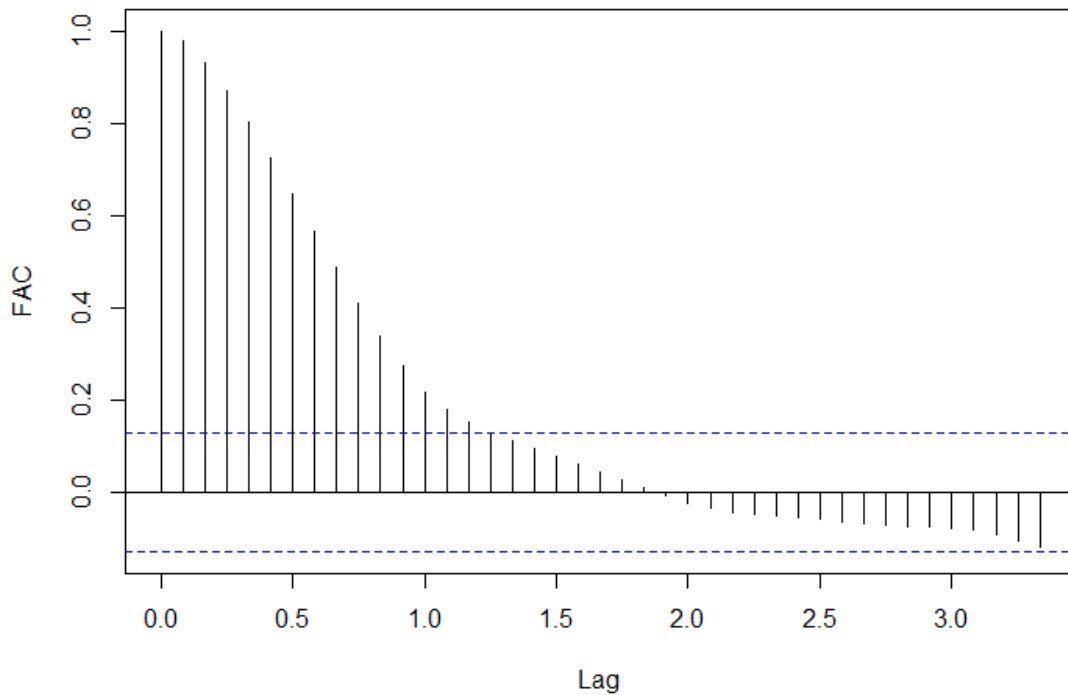


Figura 4.13: Função de Autocorrelação da série (Cenário B)

A Figura 4.14 representa a Distribuição Média, sendo possível identificar, da mesma forma que no cenário A, que um modelo aditivo é adequado a ser adotado para estimação. Além disso, verifica-se que não é necessário o uso de transformações, uma vez que os dados comportam-se de forma aleatória em relação a amplitude e a média.

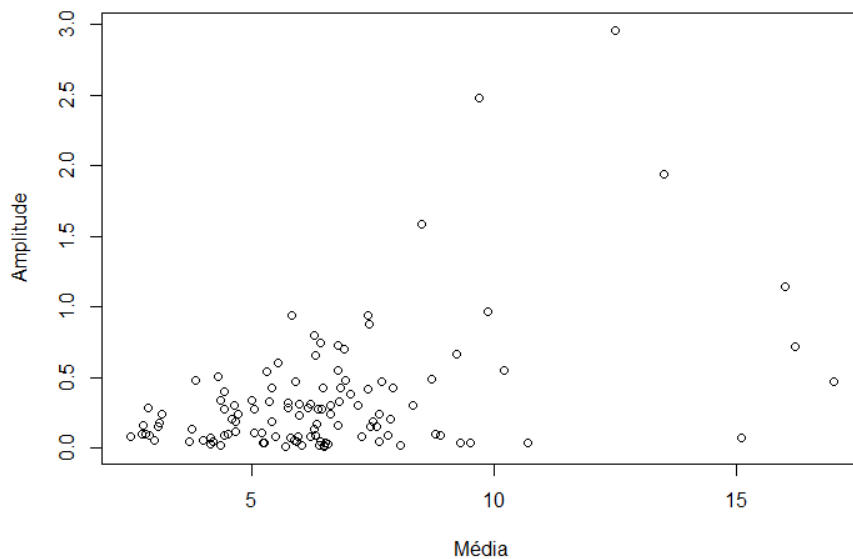


Figura 4.14: Distribuição Média pela Amplitude de grupos de observações da série (Cenário B)

A Tabela 4.13 apresenta o resultado dos testes para tendência e por meio deles, concluiu-se que a hipótese nula foi rejeitada ao nível de significância de 5% para todos eles. Sendo assim, deve-se tomar uma diferença na série a fim de eliminar a tendência identificada.

Tabela 4.13: Teste para verificação de tendência na série (Cenário B)

Teste	P-valor	Conclusão
Wald-Wolfowitz	$<2,2e-16$	Rejeita-se H0
Cox-Stuart	$1,468e-07$	Rejeita-se H0
Correlação de Spearman	$1,298e-06$	Rejeita-se H0

Após efetuar a diferença na série, no intuito de eliminar a tendência dos dados e torná-los estacionários, aplicou-se novamente os testes de tendência e pôde constatar a eliminação da componente. A Figura 4.15 ilustra a série com a primeira diferença. Da mesma forma que no Cenário A, os gráficos da Função de Autocorrelação e Autocorrelação Parcial da série diferenciada, representados pela Figura 4.16, apresentam um padrão de defasagens significativas, na FAC, e a existência de “lags” múltiplos de 12 significativos na FACp, reforçando a escolha pelo ajuste de um modelo do tipo SARIMA.

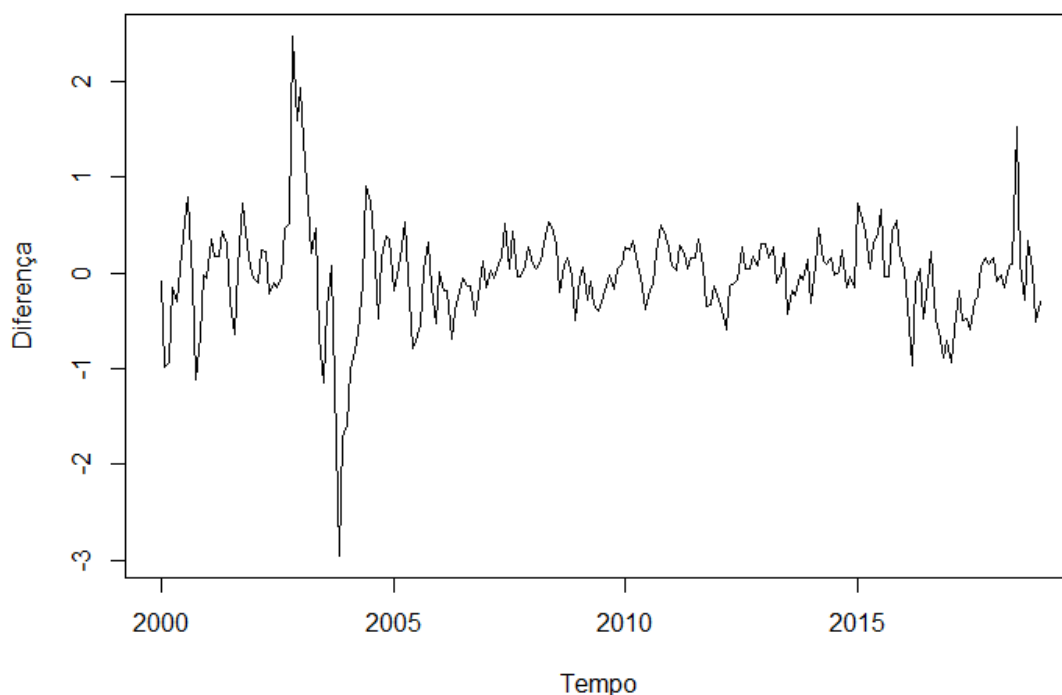


Figura 4.15: Gráfico da série original com uma diferença (Cenário B)

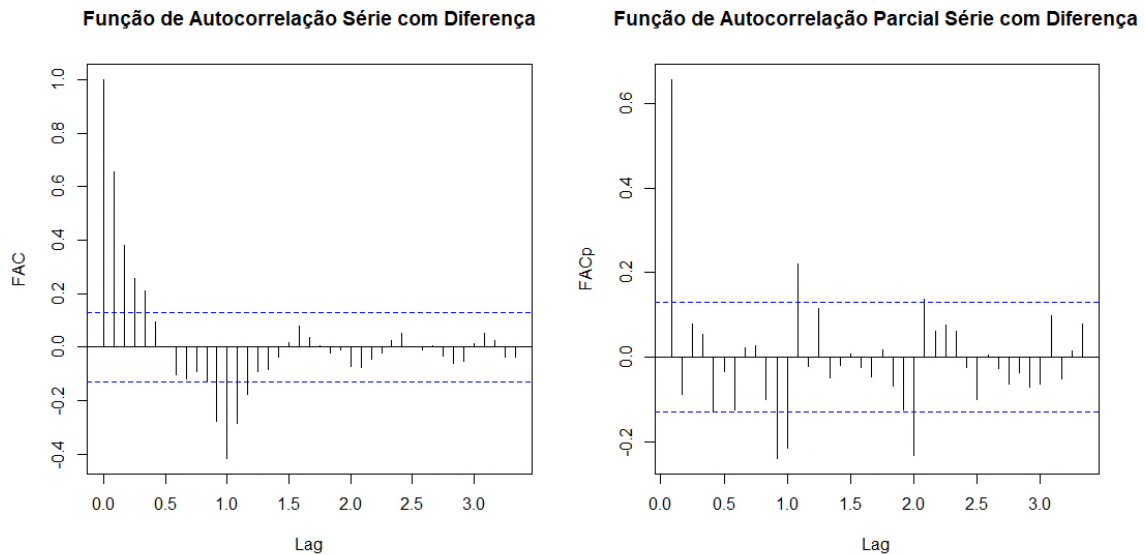


Figura 4.16: Gráficos de FAC e FACp da série diferenciada (Cenário B)

Em seguida, aplicou-se os testes de sazonalidade de Friedman e Fisher. A Tabela 4.14 apresenta os resultados obtidos para os testes, em que ambos indicam a existência da componente de sazonalidade, corroborando com a ideia de ajuste de modelos do tipo sazonais.

Tabela 4.14: Teste para verificação de sazonalidade na série (Cenário B)

Teste	P-valor	Conclusão
Friedman	$< 2,2e-16$	Rejeita-se H_0
Fisher	$2,458875e-12$	Rejeita-se H_0

Na sequência, iniciou-se a etapa de estimação do modelo. Para ajuste dos modelos, fixou-se o valor de d em 1, uma vez identificado que apenas uma diferença é suficiente para remover a tendência dos dados. Os modelos propostos foram os mesmos utilizados no “Cenário A”, representados na Tabela 4.3 e obtiveram os mesmos resultados de significância dos parâmetros.

Para seleção dos modelos, assim como no Cenário A, adotou-se como critério inicial de seleção a significância de todos os parâmetros. Selecionou-se apenas os modelos 7 (SARIMA $(1,1,0)(2,0,0)_{[12]}$), 8 (SARIMA $(1,1,0)(1,0,0)_{[12]}$) e 11 (SARIMA $(1,1,0)(3,0,0)_{[12]}$). Obteve-se para cada um dos três modelos selecionado os valores para os critérios de seleção: RMSE, MAPE, AIC e BIC, retratados na Tabela 4.15.

Tabela 4.15: Critérios de seleção dos modelos SARIMA propostos (Cenário B)

Modelo	RMSE	MAPE	AIC	BIC
Modelo 7 - SARIMA $(1,1,0)(2,0,0)_{[12]}$	0,50486	0,10428	161,97090	175,68830
Modelo 8 - SARIMA $(1,1,0)(1,0,0)_{[12]}$	0,83247	0,22412	186,61370	196,90170
Modelo 11 - SARIMA $(1,1,0)(3,0,0)_{[12]}$	0,90178	0,16533	152,39280	169,53950

O Modelo 7 (SARIMA $(1,1,0)(2,0,0)_{[12]}$) apresentou os menores valores para o RMSE e o MAPE, enquanto que o Modelo 11 (SARIMA $(1,1,0)(3,0,0)_{[12]}$) os menores valores para o AIC e

BIC. Pelo Princípio da Parcimônia e por apresentar menores valores de erros de previsão, optou-se por adotar o Modelo 7, visto que este apresenta uma quantidade menor de parâmetros. Os parâmetros estimados são apresentados na Tabela 4.16.

Tabela 4.16: Parâmetros estimados do modelo SARIMA (1,1,0)(2,0,0)_[12]

Modelo	Parâmetro	Estimativa	Desvio-Padrão
Modelo 7 - SARIMA (1,1,0)(2,0,0) _[12]	ϕ_1	0,6628	0,0495
	Φ_1	-0,6205	0,0672
	Φ_2	-0,3573	0,0660

Avaliando as funções de autocorrelação (FAC) e autocorrelação parcial (FACp) residuais, representados na Figura 4.17, para validação da adequabilidade do modelo, é possível observar a existência de um “lag” significativo na FAC e dois “lags” significativos na FACp, indicando que os resíduos são não correlacionados. Para o teste de Box-Pierce, obteve-se p-valor de 0,8025, constatando que os resíduos são um processo de ruído branco, fato confirmado pelo teste do periodograma acumulado, apresentado na Figura 4.18.

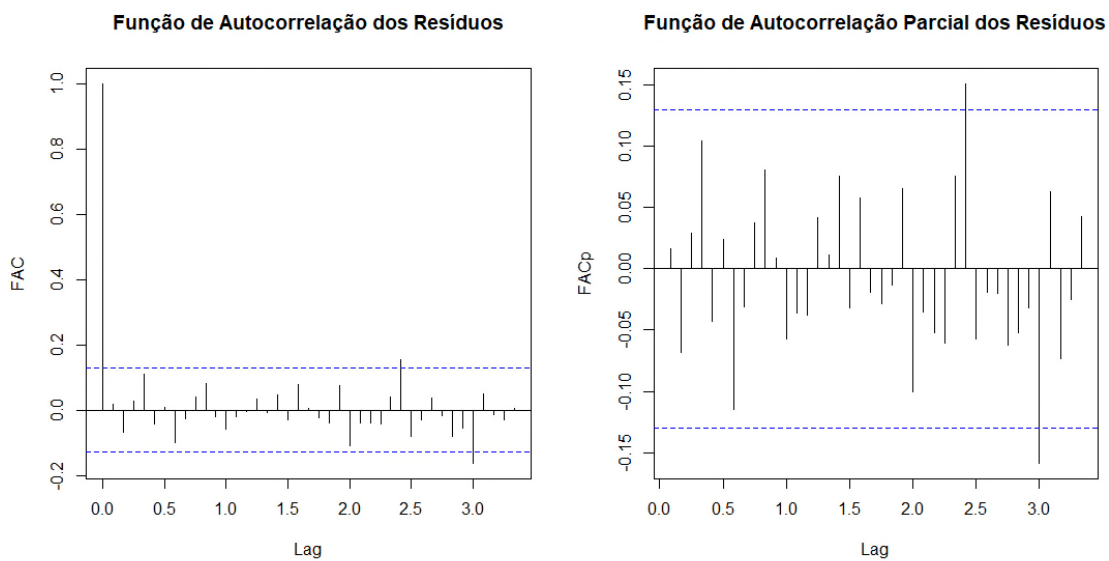


Figura 4.17: Gráficos de FAC e FACp dos Resíduos (Cenário B)

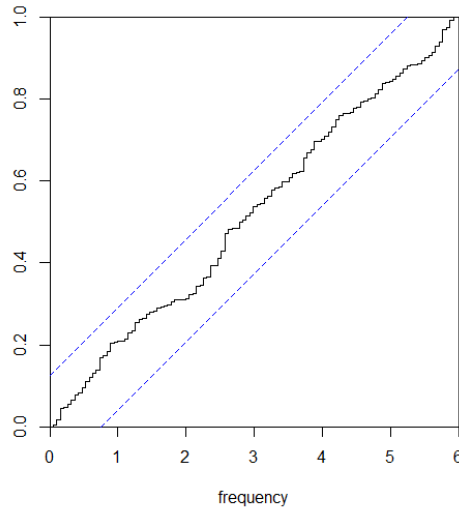


Figura 4.18: Periodograma Acumulado (Cenário B)

Validado o modelo proposto, obteve-se as previsões para o período de Janeiro/2019 à Dezembro/2019. A Figura 4.19 ilustra os valores obtidos comparados aos valores observados neste mesmo período. Observa-se que os valores previstos estão próximos dos reais, indicando que o modelo ajustado é adequado para previsão do índice de inflação no Brasil. Na Tabela 4.17, é possível confirmar que os erros de previsões foram relativamente baixos, quando comparados com os resultados do “Cenário A”.

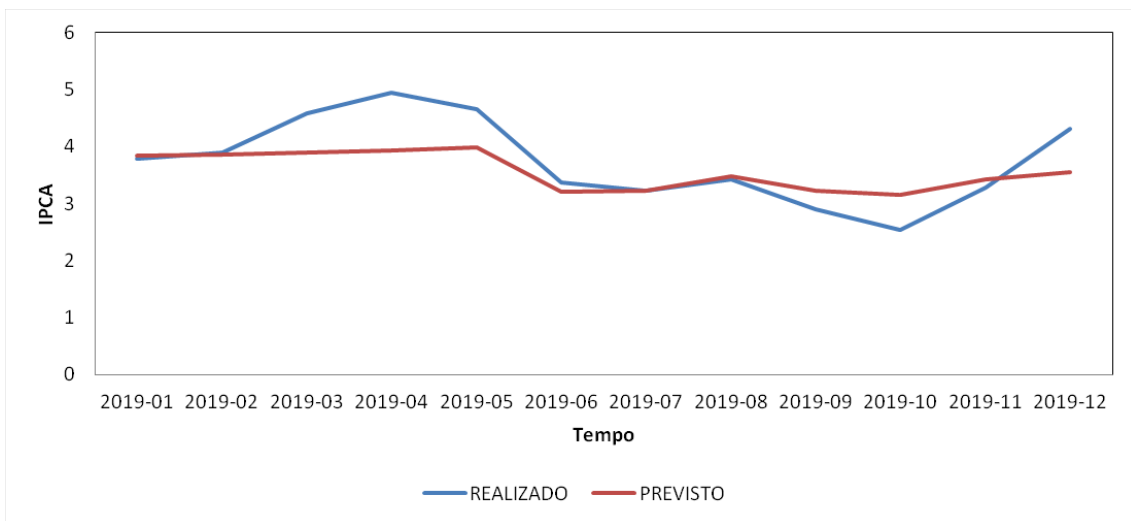


Figura 4.19: Previsões do modelo SARIMA (Cenário B)

Tabela 4.17: Previsão realizada utilizando modelo SARIMA e erro de previsão (Cenário B)

Período	Valor Real	Valor Previsto	Erro de Previsão
2019-01	3,78	3,83878	0,05878
2019-02	3,89	3,86122	0,02879
2019-03	4,58	3,89532	0,68468
2019-04	4,94	3,93255	1,00745
2019-05	4,66	3,98354	0,67646
2019-06	3,37	3,20980	0,16020
2019-07	3,22	3,23189	0,01189
2019-08	3,43	3,48413	0,05413
2019-09	2,89	3,23330	0,34330
2019-10	2,54	3,15005	0,61005
2019-11	3,27	3,42582	0,15582
2019-12	4,31	3,55509	0,75491

4.2.2 MODELOS LSTM

LSTM SEM REPOSIÇÃO

Da mesma forma que realizado anteriormente para o “Cenário A”, foram ajustados 20 modelos LSTM Sem Reposição variando apenas a quantidade de neurônios e épocas. Obteve-se então os valores do RMSE para cada ajuste, apresentados na Tabela 4.18. O modelo que apresentou o menor valor foi o LSTM Sem Reposição com 7 Neurônios e 500 Épocas.

Tabela 4.18: Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Sem Reposição (Cenário B)

Neurônios	Épocas			
	100 Épocas	300 Épocas	500 Épocas	1000 Épocas
7	0,840167	0,874836	0,656259	0,864261
25	0,816595	0,842409	1,016555	0,996567
50	0,873526	0,935319	2,081523	0,834141
100	1,238573	1,465272	1,055164	1,450778
200	1,369546	0,861307	1,019192	1,932660

Definido o melhor modelo, obteve-se um gráfico com os valores reais e previstos pelo ajuste no período de Janeiro/2019 à Dezembro/2019, representado na Figura 4.20. As mesmas representações podem ser analisadas pela Tabela 4.19

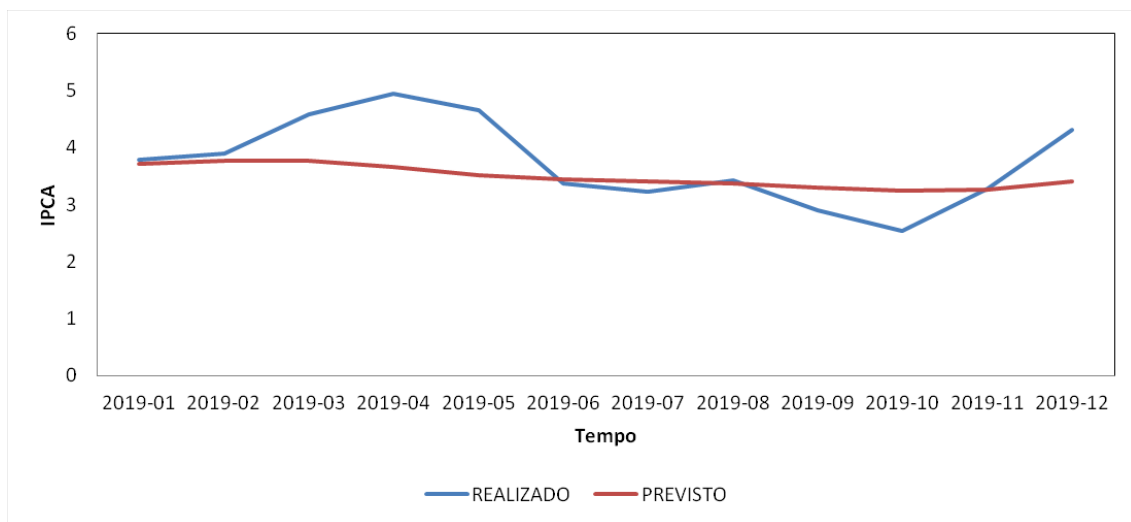


Figura 4.20: Previsões do modelo LSTM Sem Reposição (Cenário B)

Tabela 4.19: Previsão realizada utilizando modelo LSTM Sem Reposição e erro de previsão (Cenário B)

Período	Valor Real	Valor Previsto	Erro de Previsão
2019-01	3,78	3,7159	0,06409
2019-02	3,89	3,7721	0,11794
2019-03	4,58	3,7663	0,81367
2019-04	4,94	3,6517	1,28834
2019-05	4,66	3,5195	1,14051
2019-06	3,37	3,4332	0,06317
2019-07	3,22	3,4014	0,18144
2019-08	3,43	3,3781	0,05186
2019-09	2,89	3,3024	0,41241
2019-10	2,54	3,2360	0,69605
2019-11	3,27	3,2696	0,00044
2019-12	4,31	3,3971	0,91285

Apesar de apresentar um bom valor para RMSE, bem como erros de previsão relativamente baixos, é possível observar na Figura 4.20 que os valores previstos estão se comportando quase que de forma linear, levando a questionar a real qualidade do modelo ajustado, visto que pode ter ocorrido algum erro de generalização.

LSTM Com REPOSIÇÃO

Assim como no cenário anterior, ajustou-se os 20 modelos de LSTM Com Reposição, variando as quantidades de neurônios e épocas, e calculando o RMSE a fim de selecionar o melhor modelo entre os ajustados. A Tabela 4.20 apresenta os resultados obtidos.

Tabela 4.20: Valores de RMSE para testes variando Neurônios e Épocas em modelos LSTM Com Reposição (Cenário B)

Neurônios	Épocas			
	100 Épocas	300 Épocas	500 Épocas	1000 Épocas
7	0,576789	0,562368	0,622436	0,807576
25	0,597113	0,758777	0,557478	0,641260
50	0,595374	0,555143	0,548486	0,567798
100	0,492127	0,588366	0,550778	0,854994
200	0,572863	0,589755	0,848020	0,676575

O modelo que apresentou o melhor ajuste seguindo o critério estabelecido foi o LSTM Com Reposição com 100 Neurônios e 100 Épocas. Da mesma forma que encontrado no “Cenário A”, a quantidade de épocas foi 100, diferenciando apenas na quantidade de neurônios, que passou de 7 para 100. Esse fator pode ser justificado pelo fato dos dados no período de Janeiro/2019 à Dezembro/2019 apresentarem uma variabilidade menor, comparados ao período de Maio/2020 à Abril/2021, necessitando de mais iterações na camada intermediária para conseguir realizar as previsões.

A Figura 4.21 e a Tabela 4.21 apresentam detalhadamente os valores previstos comparados aos valores reais do índice de inflação no Brasil.

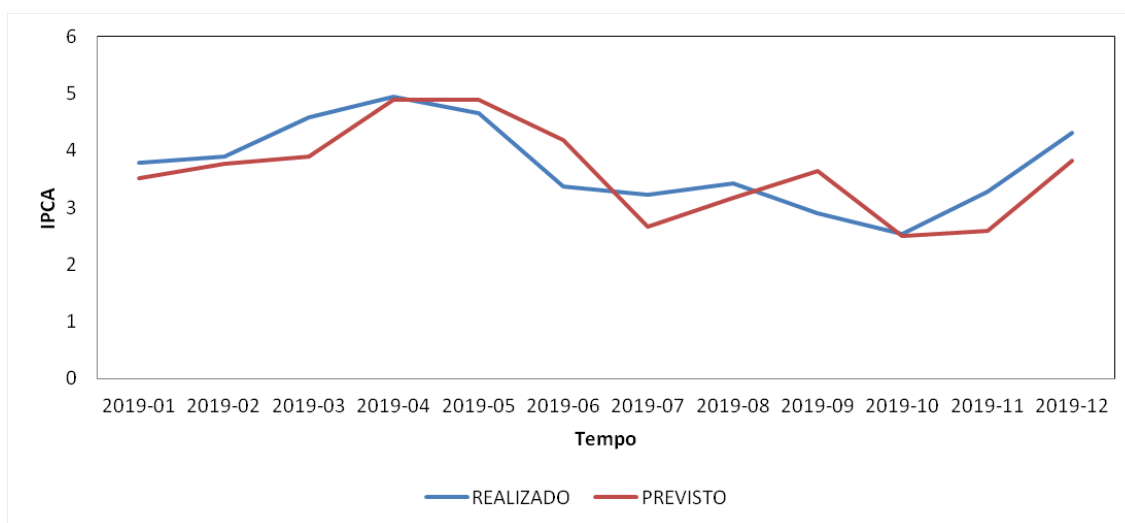


Figura 4.21: Previsões do modelo LSTM Com Reposição (Cenário B)

Tabela 4.21: Previsão realizada utilizando modelo LSTM Com Reposição e erro de previsão (Cenário B)

Período	Valor Real	Valor Previsto	Erro de Previsão
2019-01	3,78	3,50740	0,27260
2019-02	3,89	3,77320	0,11680
2019-03	4,58	3,90202	0,67798
2019-04	4,94	4,88349	0,05651
2019-05	4,66	4,88452	0,22452
2019-06	3,37	4,18050	0,81050
2019-07	3,22	2,66190	0,55810
2019-08	3,43	3,17526	0,25474
2019-09	2,89	3,64633	0,75633
2019-10	2,54	2,49550	0,04450
2019-11	3,27	2,59144	0,67856
2019-12	4,31	3,82283	0,48717

4.2.3 COMPARAÇÃO ENTRE OS MODELOS SARIMA E LSTM

Por fim, comparou-se os resultados obtidos nas técnicas aplicadas. Na Tabela 4.22 está representada o valor do RMSE obtido por cada um dos modelos selecionados.

Tabela 4.22: Valores de RMSE dos modelos selecionados de cada técnica estudada (Cenário B)

Modelo	RMSE
Modelo 7 - SARIMA(1,1,0)(2,0,0)	0,50486
LSTM Sem Reposição - 7 Neurônios e 500 Épocas	0,65626
LSTM Com Reposição - 100 Neurônios e 100 Épocas	0,49213

Na Figura 4.22, projetou-se as estimativas de todos os modelos selecionados para cada técnica comparados ao valor real da inflação no período de Janeiro/2019 à Dezembro/2019.

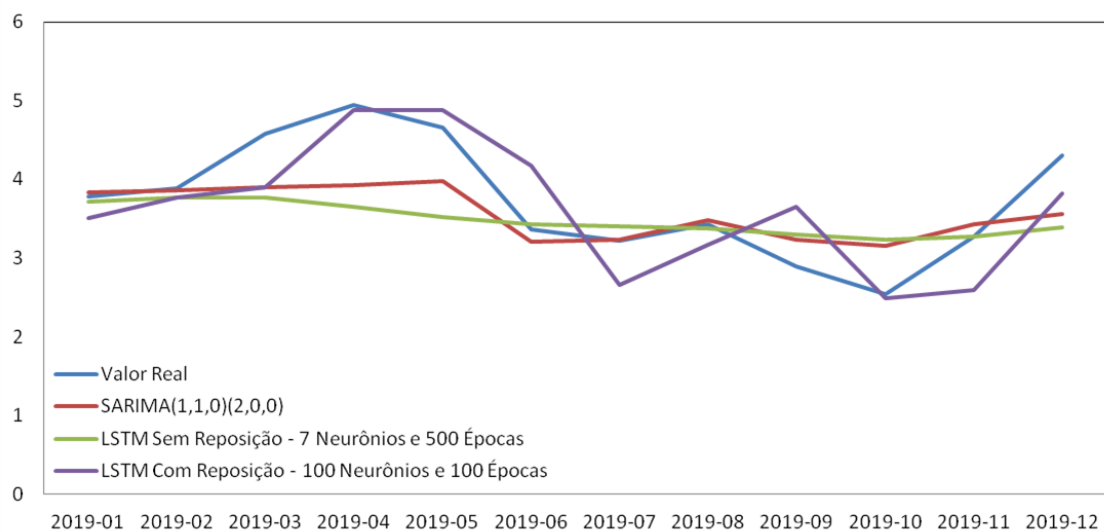


Figura 4.22: Série Original e Previsões dos modelos (Cenário B)

Seguindo o critério de seleção do menor valor para o RMSE, tem-se que o modelo que mais se aproximou dos valores observados é o LSTM Com Reposição - 100 Neurônios e 100 Épocas. Comparando os modelos SARIMA(1,1,0)(2,0,0) e LSTM Sem Reposição - 7 Neurônios e 500 Épocas, que apresentam a mesma ideia de estimação de 12 valores de uma só vez, temos que a técnica proposta por Box & Jenkins[8] obteve melhores resultados, visto que o modelo de RNN apresentou um comportamento de *overtraining*, quando a rede se especializa no conjunto de dados de treinamento por conta da grande quantidade de treinos e perde a capacidade de generalização, indicando que talvez o ajuste não tenha sido adequado.

5. CONCLUSÕES

Os modelos obtidos pelas técnicas de Box & Jenkins e redes neurais recorrentes se apresentaram úteis para descrever o comportamento do índice de inflação medido pelo IPCA no Brasil. Ao todo, foram escolhidos 6 modelos entre todos os ajustados, para as técnicas SARIMA, LSTM Sem Reposição e LSTM Com Reposição, sendo 3 para cada cenário.

No Cenário A, que analisa todo o período de Dezembro/1999 à Abril/2021, o modelo SARIMA que apresentou os melhores resultados foi o SARIMA(1,1,0)(1,0,0)_[12], e para os modelos de RNN foram o LSTM Sem Reposição com 25 neurônios e 300 épocas e o LSTM Com Reposição com 7 neurônios e 100 épocas. Dentre estes, o melhor ajuste foi este último, que apresenta comportamento de realimentação de informações, de forma a tornar as previsões mais acertivas. Apesar deste modelo ter se mostrado adequado, observou-se que fatores externos estariam afetando na estimação, em decorrência da pandemia do COVID-19. Portanto, realizou-se o ajuste de modelos em um período sem a influência deste fenômeno, o qual chamamos de *Cenário B*.

Para o Cenário B, que analisa apenas o período de Dezembro/1999 à Dezembro/2019, os melhores modelos foram: SARIMA(1,1,0)(2,0,0)_[12], LSTM Sem Reposição com 7 neurônios e 500 épocas e o LSTM Com Reposição com 100 neurônios e 100 épocas. Neste contexto, o melhor ajuste foi novamente o com reposição de informações. Porém, ao se tratar de modelos que fazem a previsão de 12 meses futuros de uma só vez, o modelo mais qualificado foi o SARIMA, visto que apresentou o segundo menor valor para o RMSE e pelo fato do modelo sem reposição ter apresentado um comportamento de overtraining, que é quando o modelo começa a perder a capacidade de generalizar os valores previstos em decorrência de um número muito grande de treinamento.

Dentre os dois cenários estudados, os melhores resultados obtidos foram para o Cenário B, visto que apresentaram valores de erros de previsão menores do que o Cenário A. Essa conclusão se dá por conta de o segundo cenário estar livre de grandes ruídos externos como é o caso da pandemia da COVID-19, que acaba por ter grande impacto nos valores realizados.

Por fim, conclui-se que quando o objetivo for estimar valores futuros a curto prazo (no máximo 3 meses), a técnica de LSTM se torna uma opção viável, tendo em vista a facilidade do ajuste dos modelos. Por outro lado, se a intenção for fazer uma previsão única a um médio prazo (12 meses), recomenda-se o uso das técnicas propostas por Box & Jenkins, tendo em vista que estes modelos são mais robustos e podem evitar erros de generalização.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Academy, D. S.: *Deep Learning Book*, 2021. <https://www.deeplearningbook.com.br/>, acessado em 22/05/2021.
- [2] Akaike, H.: *A new look at the statistical model identification*). IEEE Transactions on Automatic Control, 19(06):716–723, 1974.
- [3] Andrade, C. A.: *Análise de séries temporais aplicada ao programa de coleta seletiva no município de Uberlândia (MG)*. Monografia, 2019.
- [4] Arestis, P., Paula, L. F. de e Filho, F. F.: *A nova política monetária: uma análise do regime de metas de inflação no Brasil*. Economia e Sociedade, 18(1):1–30, 2009.
- [5] BACEN: *Banco Central do Brasil*, 2021. <https://www.bcb.gov.br/>.
- [6] Barbosa, L. M.: *Prevenção a Inflação Brasileira: Comparação Entre um Modelo Teórico e um Ateórico de Séries Temporais*. Boletim Informações Fipe, pp. 44–49, 2019.
- [7] Bonno, S. J. P.: *Previsão de Inflação Utilizando Modelos de Séries Temporais*. Dissertação de Mestrado, 2014.
- [8] Box, G. E., Jenkins, G. M. e Reinsel, G. C.: *Time Series Analysis: Forecasting and Control*, vol. 1. Wiley, 5ª ed., 2015.
- [9] Brownlee, J.: *Machine Learning Mastery*, 2021. <https://machinelearningmastery.com/category/lstm/>.
- [10] Colaboratory, G.: *Google Colaboratory*, 2021. <https://research.google.com/colaboratory/>.
- [11] Conover, W. J.: *Practical Nonparametric Statistics (Wiley Series in Probability and Statistics)*. John Wiley Sons, 3ª ed., 1980.
- [12] Dametto, R. C.: *Estudo da aplicação de redes neurais artificiais para predição de séries temporais financeiras*. Dissertação de Mestrado, 2018.
- [13] Ehlers, R. S.: *Análise de Séries Temporais*, 2009. <http://www.icmc.usp.br/ehlers/stemp/stemp.pdf>, acessado em 28/03/2021.

- [14] Foundation, T. D.: *LibreOffice Calc.*, 2021. <https://www.libreoffice.org/discover/calc/>.
- [15] Haykin, S.: *Redes neurais: princípios e prática*. Bookman Editora, 2ª ed., 2007.
- [16] Hochreiter, S. e Schmidhuber, J.: *Long Short-Term Memory*. 9(8):1735–1780, 1997.
- [17] Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O’Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. e Yasmeeen, F.: *forecast: Forecasting functions for time series and linear models*, 2021. <https://pkg.robjhyndman.com/forecast/>, R package version 8.15.
- [18] IBGE: *Instituto Brasileiro de Geografia e Estatística*, 2021. <https://www.ibge.gov.br/>.
- [19] Kopecky, K. J. e Pierce, D. A.: *Efficiency of Smooth Goodness-of-Fit Tests*. Journal of the American Statistical Association, 74(366a):393–397, 1979.
- [20] Mendonça, H. F.: *Mecanismos de transmissão monetária e a determinação da taxa de juros: uma aplicação da regra de Taylor ao caso brasileiro*. Economia e Sociedade, 10(1):65–81, 2016.
- [21] Morettin, P. A. e Toloi, C. M. C.: *Análise de séries temporais*. Blucher, 2ª ed., 2006.
- [22] Neto, A.: *Análise de Séries Temporais - Notas de Aula*, 2009. https://docs.ufpr.br/~soniaisoldi/ce001/CE017_09_08.pdf, acessado em 28/05/2021.
- [23] Olah, C.: *Understanding LSTM Networks*, 2015. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>, acessado em 22/05/2021.
- [24] OMS: *Organização Mundial de Saúde*, 2021. <https://www.who.int/es>.
- [25] Python: *Python Software Foundation*, 2021. <https://www.python.org/>.
- [26] SIDRA: *Sistema IBGE de Recuperação Automática - SIDRA*, 2021. <https://sidra.ibge.gov.br/>.
- [27] SILVA, I. N., SPATTI, D. H. e A.FLAUZINO, R.: *Redes Neurais Artificiais para engenharia e ciências aplicadas*. Artliber, 1ª ed., 2010.
- [28] Souza, R. C.: *Metodologias para a análise e previsão de séries temporais univariadas e multivariadas*. Brazilian Review of Econometrics, 1(2):77–105, 1981.
- [29] Team, R. C.: *R: A Language and Environment for Statistical Computing*, 2021. <https://www.R-project.org/>.
- [30] Vassalli, L. C.: *Aplicação de redes neurais LSTM para a previsão de curto prazo de vazão do Rio Paraíba do Sul*. Monografia, 2018.