

UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA

LUCAS LARSSON VIDIGAL

**USO DE SÉRIES TEMPORAIS E GRÁFICOS DE CONTROLE PARA
PREVISÃO E MONITORAMENTO DO NÍVEL DE ATIVIDADE DA
ECONOMIA BRASILEIRA**

SÃO PAULO
2021

LUCAS LARSSON VIDIGAL

**USO DE SÉRIES TEMPORAIS E GRÁFICOS DE CONTROLE PARA
PREVISÃO E MONITORAMENTO DO NÍVEL DE ATIVIDADE DA
ECONOMIA BRASILEIRA**

Trabalho de Formatura apresentado à
Escola Politécnica da Universidade de
São Paulo para obtenção do Diploma de
Engenheiro de Produção.

Orientadora: Profa. Dra. Linda Lee Ho

SÃO PAULO

2021

LUCAS LARSSON VIDIGAL

**USO DE SÉRIES TEMPORAIS E GRÁFICOS DE CONTROLE PARA
PREVISÃO E MONITORAMENTO DO NÍVEL DE ATIVIDADE DA
ECONOMIA BRASILEIRA**

Trabalho de Formatura apresentado à
Escola Politécnica da Universidade de
São Paulo para obtenção do Diploma de
Engenheiro de Produção.

Orientadora: Prof. Dra. Linda Lee Ho

SÃO PAULO

2021

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo-na-publicação

Vidigal, Lucas Larsson

Uso de séries temporais e gráficos de controle para previsão e monitoramento do nível de atividade da economia brasileira / L. L. Vidigal -- São Paulo, 2021.
120 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1.Análise de séries temporais 2.Modelos em séries temporais
3.Controle estatístico do processo 4.Macroeconomia I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

*A meus pais e à minha namorada, por todo o
apoio e suporte.*

AGRADECIMENTOS

Agradeço, primeiramente, à Profa. Dra. Linda Lee Ho, que ao longo dos dois últimos semestres me orientou na realização do trabalho de formatura de maneira muito próxima, esclarecendo dúvidas, dando conselhos e acompanhando todo o processo de confecção do trabalho.

Gostaria de agradecer também aos meus colegas de trabalho e gestores com quem eu tive a oportunidade de conviver nos estágios que realizei durante a graduação, que me proporcionaram ensinamentos e conselhos para o meu desenvolvimento profissional e pessoal.

Aos meus amigos da Escola Politécnica, em especial Artur, João, Pedro, Ricardo e Rodrigo, que me acompanharam ao longo de todos esses anos de graduação, por todas as experiências que tivemos dentro e fora da faculdade e pelo companheirismo. Agradeço ainda aos amigos que fiz durante meu intercâmbio na UCLouvain, com quem eu pude compartilhar meus dias fora do país.

À minha família toda, que esteve sempre presente em todos os momentos da minha vida e me deu todo o apoio para eu ser a pessoa que sou hoje. Agradeço especialmente a meus pais, Francisco e Ingrid, por todo o esforço que colocaram na minha formação e por toda a educação que me proporcionaram desde que nasci; e a meus irmãos, Ana Luiza e Rodrigo, pelo companheirismo ao longo de todos os anos, e Olívia, por trazer mais alegria à nossa família.

Por fim, gostaria de agradecer à minha companheira de todas as horas, Helena, que esteve comigo nos bons e nos maus momentos dos últimos anos, sempre me apoiando, incentivando e acreditando em mim. Agradeço por todas as experiências que tivemos a oportunidade de vivenciar juntos ao longo desse período e por todas que ainda estão por vir.

RESUMO

Este trabalho de formatura propõe o desenvolvimento de uma metodologia para realizar a previsão e o monitoramento do nível de atividade da economia brasileira, através do ajuste de um modelo paramétrico de séries temporais sobre a série do IBC-Br e da utilização de CEP. Com este modelo, busca-se criar uma ferramenta quantitativa que auxilie investidores no processo de tomada de decisão para alocação de seus ativos, podendo ser combinada com análises qualitativas de mercado. Para tal, inicialmente, decidiu-se por utilizar um modelo da classe SARIMA para ajustar a série do IBC-Br. Procurou-se obter um modelo que, ao mesmo tempo, apresentasse boa aderência à série e fosse parcimonioso. Sua avaliação foi feita com base nas previsões realizadas para os dados de teste da série, que não foram utilizados para a modelagem, a partir de dois critérios: RMSE e aderência aos intervalos de confiança. Em seguida, para monitoramento do IBC-Br, foi empregado um gráfico de controle com memória, do tipo EWMA, com um fator λ de 0,6. Os limites de controle foram determinados através de simulações visando obter um ARL alvo para a série de 36. Para avaliar a capacidade do gráfico em sinalizar processos fora de controle, foram realizadas simulações com maior variabilidade da série modelada. Por fim, realizou-se a avaliação do gráfico EWMA, aplicando-o à série completa do IBC-Br e verificando se ele conseguiu identificar de forma assertiva períodos durante os quais o processo estava fora de controle. A ferramenta desenvolvida apresentou resultados satisfatórios tanto para a previsão quanto para o monitoramento da variável.

Palavras-chave: séries temporais; modelos paramétricos; sazonalidade; gráficos de controle; monitoramento de processos; economia brasileira.

ABSTRACT

This paper proposes the development of a methodology to forecast and monitor the level of activity of the Brazilian economy, through the fitting of a parametric model of time series on the IBC-Br data and the use of statistical process control. With this model, we seek to create a quantitative tool to assist investors in their asset allocation decision making process, which can be combined with qualitative market analyses. To this end, it was initially decided to use a SARIMA class model to adjust the IBC-Br series. We tried to obtain a model that, at the same time, presented good adherence to the series and was parsimonious. Its evaluation was based on the forecasts made for the test data of the series, which were not used for modeling, based on two criteria: RMSE and adherence to confidence intervals. Then, for monitoring the IBC-Br, a control chart with memory was employed, of the EWMA type, with a λ factor of 0.6. The control limits were determined through simulations aimed at obtaining a target ARL for the series of 36. To evaluate the chart's ability to signal out of control processes, simulations were performed with greater variability of the modeled series. Finally, the EWMA chart was evaluated by applying it to the complete IBC-Br series and verifying if it could assertively identify periods during which the process was out of control. The tool developed presented satisfactory results both for forecasting and monitoring the variable.

Keywords: time series; parametric models; seasonality; control charts; process monitoring; Brazilian economy.

LISTA DE FIGURAS

Figura 1 - Variações interanuais do PIB do IBC-Br	20
Figura 2 - Variações anuais do PIB e do IBC-Br	20
Figura 3 - Gráfico QQ da série de retornos diários do Ibovespa.....	31
Figura 4 - Série temporal de passageiros internacionais no transporte aéreo	38
Figura 5 - Exemplo de gráfico de controle	43
Figura 6 - Fluxograma com diretrizes para a seleção do gráfico de controle	50
Figura 7 - Número-índice do IPCA	54
Figura 8 - Variação mensal do IPCA	55
Figura 9 - FAC de um processo AR(2) puro	57
Figura 10 - FACP de um processo AR(2) puro	57
Figura 11 - Gráfico com os p-valores de cada <i>lag</i> do teste Ljung-Box para o modelo ajustado ao processo AR(2)	62
Figura 12 - Fluxograma do algoritmo para cálculo dos limites de controle	65
Figura 13 - Fluxograma do algoritmo para cálculo dos ARLs fora de controle	67
Figura 14 - Série temporal do IBC-Br	71
Figura 15 - Série de retornos mensais do IBC-Br.....	72
Figura 16 - Decomposição de R_t	73
Figura 17 - Série de retornos do IBC-Br diferenciada sazonalmente	74
Figura 18 - Decomposição de D_t	74
Figura 19 - Gráfico QQ da série D_t	75
Figura 20 - Gráfico de autocorrelação da série D_t	76
Figura 21 - Gráfico de autocorrelação parcial da série D_t	76
Figura 22 - Resíduos do modelo selecionado	79
Figura 23 - Fac dos resíduos do modelo	80
Figura 24 - Facp dos resíduos do modelo	80
Figura 25 - Representação gráfica do teste de Ljung-Box.....	81
Figura 26 - Gráfico QQ dos resíduos	82
Figura 27 - ARL em função de L	83
Figura 28 - ARL em função de k	84
Figura 29 - Série de treinamento vs. valores ajustados pelo modelo.....	85

Figura 30 - Série de teste vs. previsões feitas pelo modelo com intervalo de confiança de 90%	86
Figura 31 - Série completa vs. previsões feitas pelo modelo	87
Figura 32 - Gráfico de controle EWMA da estatística Z_t	89
Figura 33 – Série estacionária D_t com destaque para os pontos fora de controle ...	90
Figura 34 - Série de retornos R_t com destaque para os pontos fora de controle	91

LISTA DE TABELAS

Tabela 1 - AIC dos cinco modelos mais bem classificados	77
Tabela 2 - Teste de coeficientes do modelo 3	78
Tabela 3 - Teste de coeficientes do modelo 3 reduzido	78

LISTA DE ABREVIATURAS E SIGLAS

ADF	<i>Augmented Dickey-Fuller</i>
API	<i>Aplication Programming Interface</i>
AR	Modelo Autorregressivo
ARL	<i>Average Run Length</i>
ARMA	Modelo Autorregressivo de Médias Móveis
ARIMA	Modelo Autorregressivo Integrado de Médias Móveis
CEP	Controle Estatístico de Processos
CUSUM	<i>Cumulative Sum</i>
EMQ	Estimador de Mínimos Quadrados
EMV	Estimador de Máxima Verossimilhança
EWMA	<i>Exponentially Weighted Moving Average</i>
fac	Função de Autocorrelação
facp	Função de Autocorrelação Parcial
facv	Função de Autocovariância
IBC-Br	Índice de Atividade Econômica do Banco Central
IBGE	Instituto Brasileiro de Geografia e Estatística
Ibovespa	Índice da Bolsa de Valores de São Paulo
IPCA	Índice Nacional de Preços ao Consumidor Amplo
LC	Linha Central
LIC	Limite Inferior de Controle
LSC	Limite Superior de Controle
MA	Modelo de Médias Móveis
PIB	Produto Interno Bruto
RMSE	<i>Root Mean Square Error</i>
SARIMA	Modelo Autorregressivo Integrado de Médias Móveis Sazonal

SUMÁRIO

1. INTRODUÇÃO	19
1.1. Contexto	19
1.2. Definição do problema e relevância do tema.....	21
1.3. Objetivo	22
1.4. <i>Softwares</i> utilizados	23
1.5. Estrutura do trabalho.....	23
2. REVISÃO BIBLIOGRÁFICA SOBRE SÉRIES TEMPORAIS.....	25
2.1. Introdução às séries temporais	25
2.2. Processos estocásticos	26
2.3. Estacionariedade	27
2.4. Normalidade	29
2.5. Independência.....	31
2.5.1. Função de Autocovariância	32
2.5.2. Função de Autocorrelação.....	32
2.5.3. Função de Autocorrelação parcial	33
2.6. Modelos paramétricos	34
2.6.1. Modelos autorregressivos (AR)	35
2.6.2. Modelos de médias móveis (MA).....	35
2.6.3. Modelos autorregressivos de médias móveis (ARMA).....	36
2.6.4. Modelos autorregressivos integrados de médias móveis (ARIMA)	37
2.6.5. Modelos ARIMA sazonais (SARIMA).....	38
3. REVISÃO BIBLIOGRÁFICA SOBRE CONTROLE ESTATÍSTICO DE PROCESSO.....	41
3.1. Introdução ao CEP	41
3.2. Gráficos de controle	42
3.3. Tipos de gráficos de controle	45
3.3.1. Shewhart	46
3.3.2. CUSUM	47

3.3.3. EWMA	48
3.4. Escolha do gráfico de controle	50
4. METODOLOGIA.....	53
4.1. Coleta e ajuste dos dados	53
4.2. Identificação do modelo.....	55
4.3. Estimação e seleção do modelo.....	58
4.4. Diagnóstico do modelo	60
4.5. Determinação dos limites de controle	63
4.6. Obtenção dos ARL fora de controle.....	65
4.7. Avaliação dos resultados	67
5. APLICAÇÃO EM CASO REAL	71
5.1. Coleta e ajuste dos dados	71
5.2. Identificação do modelo.....	75
5.3. Estimação e seleção do modelo.....	77
5.4. Diagnóstico do modelo	79
5.5. Determinação dos limites de controle	82
5.6. Obtenção dos ARL fora de controle.....	84
5.7. Avaliação dos resultados	85
5.7.1. Previsões dos valores futuros	85
5.7.2. Gráfico de controle.....	88
6. CONCLUSÃO.....	95
REFERÊNCIAS BIBLIOGRÁFICAS	99
APÊNDICE A – CÓDIGO PARA AJUSTE DO MODELO À SÉRIE DO IBC-BR E REALIZAÇÃO DE PREVISÕES	101
APÊNDICE B – CÓDIGO PARA DETERMINAÇÃO DOS LIMITES DE CONTROLE DO GRÁFICO	107
APÊNDICE C – CÓDIGO PARA SIMULAÇÃO DOS ARLS DO PROCESSO FORA DE CONTROLE	111
APÊNDICE D – CÓDIGO PARA APLICAÇÃO DO GRÁFICO DE CONTROLE À SÉRIE DO IBC-BR.....	113
ANEXO A – SÉRIE ORIGINAL DO IBC-BR	119

1. INTRODUÇÃO

O capítulo introdutório deste trabalho de formatura tem como principal objetivo trazer o contexto no qual o trabalho está inserido, através da definição do tema que será tratado nele, da determinação de sua relevância e qual o objetivo buscado ao longo de seu desenvolvimento. Além disso, serão explicitados os *softwares* utilizados para dar suporte ao trabalho e sua estrutura, explicando como o tema será desenvolvido ao longo dos capítulos.

1.1. Contexto

O presente trabalho de formatura abrange, majoritariamente, duas áreas do conhecimento: economia e estatística. A utilização de conceitos de ambas as áreas em conjunto pode trazer importantes resultados financeiros a empresas e pessoas físicas, dado que através da utilização de ferramentas estatísticas é possível realizar análises importantes para basear a tomada de decisão com relação à alocação de recursos em ativos financeiros – como ações, títulos públicos e fundos de investimentos – e não financeiros – como imóveis.

O autor deste trabalho, nos últimos anos de graduação, realizou estágio em duas gestoras de investimentos distintas, atuando em ambas na área de pesquisa macroeconômica. Durante esse período, observou-se como é importante o acompanhamento e previsão das principais variáveis relacionadas ao desempenho da economia brasileira, dado que elas se correlacionam diretamente com diversas classes de ativos do país.

Nesse contexto, um dos principais indicadores acompanhados por gestoras de fundos de investimento é o Índice de Atividade Econômica do Banco Central (IBC-Br), que tem como principal objetivo mensurar a evolução da atividade econômica do Brasil, contribuindo, assim, para a elaboração da estratégia de política monetária e de estimativas do mercado para o Produto Interno Bruto (PIB) e para a taxa Selic, que é a taxa básica de juros da economia brasileira.

Este indicador, divulgado pelo Banco Central desde 2003, tem periodicidade mensal e utiliza como base de cálculo variáveis que buscam mensurar o desempenho de setores diversos da economia, como a agropecuária, a indústria e os serviços. Assim, sendo um indicador agregado de atividade, sua taxa de crescimento acaba

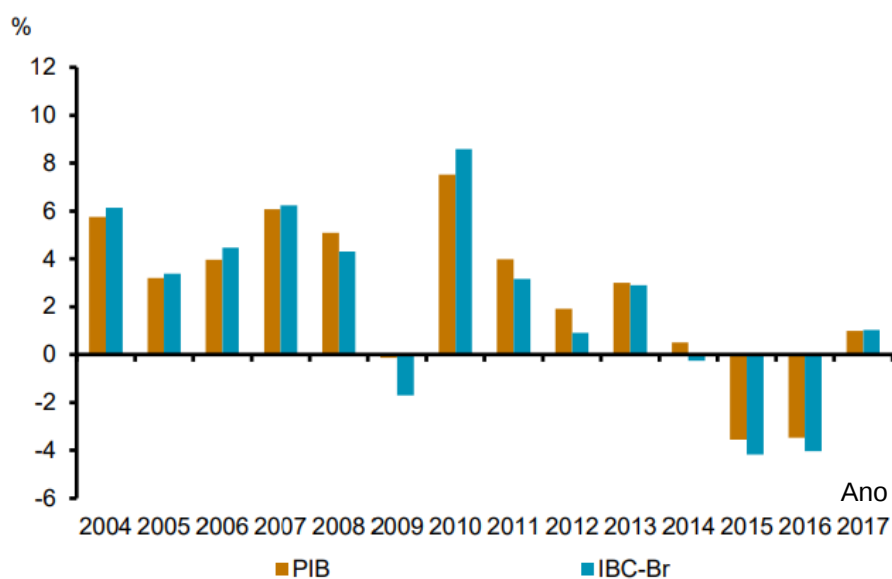
sendo muito próxima à do PIB – principalmente para horizontes mais longos –, mesmo que este se utilize de uma metodologia de cálculo bem distinta. Diferentemente do IBC-Br, o PIB é divulgado trimestralmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE). As Figuras 1 e 2 evidenciam essa semelhança entre a taxa de crescimento de ambos os índices. Na Figura 1, é trazido a variação interanual para o PIB e para o IBC-Br, ou seja, o comparativo da variável com seu valor no mesmo período do ano anterior. Já Figura 2 mostra a variação anual de cada uma delas.

Figura 1 - Variações interanuais do PIB do IBC-Br



Fonte: Banco Central (2008)

Figura 2 - Variações anuais do PIB e do IBC-Br



Fonte: Banco Central (2008).

Como é possível notar, ambos os indicadores costumam variar de maneira bem semelhante, especialmente se considerada a variação anual.

De acordo com o próprio Banco Central (2008), a criação do IBC-Br foi motivada “pela inexistência de indicador agregado de atividade econômica de frequência mensal que permitisse sintetizar e avaliar, em maior frequência, o estudo da economia, em contexto de decisões de política monetária”. Além de permitir um acompanhamento mensal da economia brasileira, o IBC-Br tem como outra vantagem sobre o PIB a rapidez de divulgação. Ao passo que o índice do Banco Central costuma ser divulgado 45 dias após o fechamento do período ao qual se refere, o PIB costuma sair aproximadamente 60 dias depois.

Dentro das gestoras de investimento, diversas das análises realizadas sobre as variáveis acompanhadas são feitas através de ferramentas estatísticas. No presente trabalho, serão abordados conceitos relacionados a dois temas principais da estatística, estudados pelo autor ao longo dos anos de graduação: séries temporais e controle estatístico de processos (CEP). Nas próximas seções deste capítulo, será especificado como estes conceitos serão utilizados para o desenvolvimento de um algoritmo para o acompanhamento do IBC-Br e quais são os principais objetivos de seu desenvolvimento.

1.2. Definição do problema e relevância do tema

Cada vez mais, o mercado financeiro brasileiro vai se tornando uma opção de investimento da população brasileira, com o aumento do número de investidores físicos na Bolsa de Valores de São Paulo e o lançamento de diversos tipos de fundos de investimento, com estratégias variadas. Com isso, as opções de investimento para pessoas físicas vão se tornando cada vez mais abrangentes. Portanto, para se manter competitivo neste mercado, é necessário inovar constantemente, de maneira a conseguir oferecer produtos cada vez melhores e com rentabilidade acima da dos concorrentes.

Nesse ambiente competitivo do mercado de gestoras de investimentos, a tomada de decisão pautada em critérios estatísticos e objetivos acabou ganhando muito espaço, com a abertura e o crescimento, inclusive, de diversos fundos cuja alocação dos ativos é pautada apenas em algoritmos.

Assim, o desenvolvimento de algoritmos baseados em modelos estatísticos e matemáticos para a previsão e monitoramento de índices econômicos e financeiros pode se tornar uma ferramenta eficiente na gestão de ativos financeiros, dado que as análises produzidas por esses algoritmos são totalmente baseadas em dados e

reduzem o impacto que vieses subjetivos podem ter no processo de tomada de decisão. Essas ferramentas, aliadas à análise qualitativa do mercado baseada no conhecimento do gestor de portfólio, podem trazer resultados interessantes para a competitividade de um fundo de investimento.

Dessa maneira, o presente trabalho se propõe a criar um modelo quantitativo para a previsão e monitoramento do IBC-Br, um dos mais importantes índices que medem o desempenho da economia brasileira, conforme explicitado na Seção 1.1. Analisar de maneira assertiva este índice pode trazer importantes resultados financeiros para gestores de fundos de investimento e investidores físicos, dado que grande parte dos ativos financeiros e não financeiros são afetados diretamente pelo nível de atividade econômica do país.

1.3. Objetivo

O principal objetivo do presente trabalho é aplicar, na prática, os conceitos de séries temporais e CEP aprendidos ao longo do curso de Engenharia de Produção e dos estágios realizados pelo autor de forma a desenvolver um modelo quantitativo para a previsão e monitoramento do nível de atividade da economia brasileira, através dos dados históricos do IBC-Br.

Para tal, o primeiro passo que se faz necessário é a modelagem da variável escolhida. Essa modelagem será realizada através de modelos de séries temporais da classe SARIMA, que será especificada ao longo do Capítulo 2. A partir do modelo selecionado para o IBC-Br, será possível cumprir a primeira parte do objetivo do presente trabalho: a realização de previsões para os valores futuros do índice.

É importante ressaltar que as previsões realizadas atualmente para o IBC-Br costumam ser mais voltadas ao curto prazo, sendo baseadas em outros indicadores de mesma periodicidade, como a Pesquisa Mensal de Serviços e a Pesquisa Industrial Mensal, ambos do IBGE. A partir da modelagem por séries temporais, com a utilização de dados históricos, será possível realizar a previsão para horizontes temporais mais distantes.

Em seguida, serão aplicadas ferramentas de CEP sobre este modelo, de forma a se obter um método para controlar a variabilidade da série, que indique situações nas quais a economia encontra-se acima ou abaixo do esperado de acordo com seu histórico. Conforme será mencionado mais à frente, a ferramenta escolhida para tal

finalidade foi o gráfico de controle, por ser a ferramenta mais robusta do CEP e por permitir uma integração interessante com modelos de séries temporais.

A partir disso, espera-se que seja possível fornecer uma ferramenta quantitativa com bom desempenho prático, que possa ser utilizada em conjunto com análises qualitativas do mercado, de maneira a basear a tomada de decisão para alocação de investimentos por parte de gestores de fundos e investidores físicos e auxiliar nas projeções para outros indicadores financeiros.

1.4. Softwares utilizados

De modo a obter os dados necessários para o trabalho, desenvolver os algoritmos propostos e realizar as análises dos resultados, foram utilizados dois principais *softwares* ao longo do desenvolvimento do presente trabalho de formatura: Microsoft Excel e RStudio.

O Microsoft Excel foi utilizado principalmente para o manuseio de algumas das séries temporais utilizadas no trabalho, além da elaboração de gráficos e tabelas que serão exibidos ao longo dos próximos capítulos.

Já o RStudio é um *software* de interface gráfica para a linguagem de programação R, largamente utilizado por estatísticos e cientistas de dados para o desenvolvimento de *softwares* estatísticos e análises de dados. Este programa foi utilizado para a elaboração dos códigos voltados ao desenvolvimento dos algoritmos e da maior parte dos gráficos para a visualização dos dados e dos resultados. Dentro do RStudio, foram acessadas funções de diversas bibliotecas externas, como “tseries” e “sarima”, que permitiram a realização das tarefas necessárias.

1.5. Estrutura do trabalho

Além do primeiro capítulo, de introdução, o presente trabalho está estruturado através de outros cinco capítulos, os quais são apresentados a seguir em conjunto com uma breve descrição:

- Capítulo 2 – Revisão bibliográfica sobre séries temporais: apresenta os principais conceitos necessários para o desenvolvimento do trabalho relacionados a séries temporais, como a estacionariedade, as funções de autocorrelação e autocorrelação parcial e as classes de modelo a serem abordadas;

- Capítulo 3 – Revisão bibliográfica sobre controle estatístico de processos: apresenta o CEP e suas ferramentas, e estuda mais a fundo os gráficos de controle e seus principais tipos, dado que é a ferramenta que será utilizada para o monitoramento da série do IBC-Br;
- Capítulo 4 – Metodologia: detalha a metodologia proposta para a solução do problema apresentado na Seção 1.2, assim como os meios através dos quais os resultados serão avaliados;
- Capítulo 5 – Aplicação em caso real: apresenta a aplicação da metodologia proposta no Capítulo 4 à série temporal do IBC-Br, assim como os resultados obtidos a partir disso;
- Capítulo 6 – Conclusão: discute os resultados obtidos pela aplicação da metodologia ao caso real e avalia o desempenho do modelo, além de destacar as principais dificuldades encontradas ao longo do desenvolvimento do trabalho e sugestões para estudos futuros.

Por fim, serão apresentados, nos apêndices e anexos, os códigos elaborados ao longo do trabalho e os dados utilizados para tal.

2. REVISÃO BIBLIOGRÁFICA SOBRE SÉRIES TEMPORAIS

O objetivo deste capítulo é explicitar a base teórica relacionada às séries temporais, um dos focos do presente trabalho, que servirá de base para o seu desenvolvimento e para a aplicação no caso real, a ser demonstrada no Capítulo 5. Assim, no Capítulo 2 será feita uma breve revisão sobre as séries temporais e suas principais propriedades, os modelos paramétricos aplicáveis a elas e o conceito de sazonalidade, que será essencial para o desenvolvimento do modelo para o IBC-Br.

2.1. Introdução às séries temporais

De acordo com Morettin e Toloi (2006), séries temporais são quaisquer conjuntos de observações ordenadas no tempo. Box, Jenkins, Reinsel e Ljung (2016), similarmente, definem séries temporais como uma sequência de observações tomadas sequencialmente no tempo. Exemplos de séries temporais podem ser encontrados nos mais diversos campos, como economia, engenharia, ciências naturais e administração de empresas. Alguns exemplos são explicitados a seguir:

- i. Valores diários de fechamento do Índice da Bolsa de Valores de São Paulo (Ibovespa);
- ii. Número semanal de acidentes de trânsito no Brasil;
- iii. Valores mensais de indicadores econômicos, como o IBC-Br;
- iv. Resultado anual de uma empresa do setor de varejo;
- v. Temperatura do ar na cidade de Manaus.

Nos exemplos acima, podemos identificar os quatro primeiros como séries discretas com relação ao tempo, enquanto o último pode ser caracterizado como uma série contínua no tempo. As séries discretas são obtidas através da amostragem de um fenômeno com uma periodicidade definida, realizada em intervalos de tempos iguais, enquanto as contínuas são compostas por observações contínuas de um fenômeno através do tempo. Na análise de séries temporais discretas no tempo, são utilizados modelos paramétricos, que são compostos por um número finito de parâmetros, ao passo que as séries contínuas podem ser modeladas por modelos não-paramétricos. O presente trabalho, por buscar a modelagem de uma série temporal discreta, com periodicidade mensal, apresentará os detalhes apenas do primeiro tipo de modelo.

Ainda de acordo com Box, Jenkins, Reinsel e Ljung (2016), uma característica intrínseca a séries temporais é que, usualmente, observações adjacentes têm dependência entre si, o que traz uma gama de aplicações a modelos que conseguem identificá-la, sendo este o principal objetivo da análise de séries temporais. Dessa maneira, os modelos paramétricos que serão estudados nas próximas seções buscam identificar justamente essa dependência. Morettin e Tolo (2006) definem quatro principais objetivos para a análise de séries temporais:

- a) Investigação do mecanismo gerador da série temporal;
- b) Previsão de valores futuros da série;
- c) Descrição do comportamento da série;
- d) Busca por periodicidades relevantes nos dados.

De todo modo, para atingir qualquer um desses objetivos, é necessária a construção de modelos probabilísticos ou estocásticos no domínio do tempo ou das frequências. No caso das séries temporais discretas com relação ao tempo, os modelos paramétricos são constituídos no primeiro domínio mencionado.

2.2. Processos estocásticos

Processos estocásticos são sequências de variáveis aleatórias que representam a evolução de um processo ao longo do tempo, sendo controlados por leis probabilísticas. Os modelos para descrição de séries temporais, objeto deste trabalho, são processos estocásticos. Morettin e Tolo (2006) estabelecem a seguinte definição para este tipo de processo:

Definição: Seja T um conjunto arbitrário. Um processo estocástico é uma família $Z = \{Z(t), t \in T\}$, tal que, para cada $t \in T$, $Z(t)$ é uma variável aleatória.

Para a classe de processos que será estudada no presente trabalho, nos interessará apenas os momentos de primeira e segunda ordem dos processos estocásticos, também chamados de média e autocovariância, respectivamente.

A função média de Z é dada por:

$$\mu(t) = E\{Z(t)\} = \int_{-\infty}^{\infty} zf(z; t)dz, \quad (2.1)$$

onde $f(z; t)$ é a função distribuição de probabilidade de $Z(t)$.

Já a função autocovariância (facv) de Z é dada por:

$$\gamma(t_\tau, t_{\tau+h}) = \text{Cov}\{Z(t_\tau), Z(t_{\tau+h})\}$$

$$= E\{Z(t_\tau)Z(t_{\tau+h})\} - E\{Z(t_\tau)\}E\{Z(t_{\tau+h})\}, \quad \tau \in T, h \in N. \quad (2.2)$$

2.3. Estacionariedade

A estacionariedade é uma das principais suposições feitas a respeito de séries temporais. Uma série estacionária desenvolve-se no tempo aleatoriamente ao redor de uma média constante, o que indica certo equilíbrio. Na prática, grande parte das séries temporais não apresentam essa característica, especialmente as séries econômicas e financeiras, que costumam ter tendências. No caso de tendências lineares, por exemplo, a média das séries varia ao longo do tempo ao redor de retas com inclinação positiva ou negativa.

De acordo com Morettin e Toloi (2006), um processo Z é estacionário se ele se desenvolver no tempo de modo que a escolha de uma origem dos tempos não seja importante. Ou seja, as características do processo $Z(t)$ são as mesmas para qualquer que seja o t escolhido. Há duas formas distintas de estacionariedade: fraca (ou ampla) e forte (ou estrita).

Definição: um processo estocástico $Z = \{Z(t), t \in T\}$ diz-se estritamente estacionário se:

$$F(z_1, \dots, z_n; t_1 + \tau, \dots, t_n + \tau) = F(z_1, \dots, z_n; t_1, \dots, t_n),$$

para quaisquer $t_1, \dots, t_n, \tau$ de T .

Dessa maneira, temos que a média $\mu(t)$ e a variância $Var(t)$ do processo são constantes ao longo do tempo. Além disso, a função autocovariância $\gamma(t_i, t_j)$ do processo depende apenas da diferença $\tau = |t_i - t_j|$. Assim, as seguintes propriedades aplicam-se a estes processos:

- i. $E\{Z(t)\} = \mu(t) = \mu$, para todo $t \in T$;
- ii. $E\{Z^2(t)\} = Var(t) = \sigma^2$, para todo $t \in T$;
- iii. $\gamma(t_i, t_j) = Cov\{Z(t_i), Z(t_j)\} = \gamma(\tau)$, para $\tau = |t_i - t_j|$.

Entretanto, no presente trabalho estaremos interessados nos processos fracamente estacionários, também chamados de estacionários de segunda ordem, doravante simplesmente processos estacionários. Morettin e Toloi (2006) os definem da seguinte forma:

Definição: um processo estocástico $Z = \{Z(t), t \in T\}$ diz-se fracamente estacionário se e somente se:

- i. $E\{Z(t)\} = \mu(t) = \mu$, para todo $t \in T$;
- ii. $E\{Z^2(t)\} = Var(t) < \infty$, para todo $t \in T$;
- iii. $\gamma(t_i, t_j) = Cov\{Z(t_i, t_j)\} = \gamma(\tau)$, para $\tau = |t_i - t_j|$.

Os modelos para séries temporais que serão apresentados na Seção 2.6 – da classe SARIMA – possuem mecanismos capazes de descrever de maneira satisfatória séries não-estacionárias homogêneas, que são séries cuja média ou inclinação mudam com o decorrer do tempo. Esta classe de séries pode tornar-se estacionária através de diferenças sucessivas, método que será descrito na mesma seção.

Para verificar se uma série temporal tem comportamento estacionário, existem diversos testes estatísticos possíveis. Um dos mais utilizados atualmente é o teste de Dickey-Fuller, proposto por Dickey e Fuller (1979). Este teste consiste em determinar se existe alguma raiz unitária para o operador autorregressivo do modelo, condição que evidencia que o processo estudado não é estacionário.

Assim, para exemplificar o teste em questão, Box, Jenkins, Reinsel e Ljung (2016) consideram o seguinte processo AR(1), conceito que será abordado na Seção 2.8.1:

$$Z_t = \phi Z_{t-1} + a_t, \quad (2.3)$$

onde a_t é ruído branco com média 0 e variância σ^2 .

O ruído branco é o exemplo mais fundamental de processo estacionário, sendo uma sequência de variáveis aleatórias independentes e identicamente distribuídas. Morettin e Toloi (2006) o define da maneira a seguir:

Definição: Dizemos que $\epsilon_t, t \in Z$ é um ruído branco discreto se as variáveis aleatórias ϵ_t são não correlacionadas, isto é, $Cov\{\epsilon_t, \epsilon_s\} = 0, t \neq s$.

Temos que a média deste processo $E\{\epsilon_t\} = \mu_\epsilon$ e a variância do processo $Var\{\epsilon_t\} = \sigma_\epsilon^2$, para todo t . A notação segue o seguinte formato: $\epsilon_t \sim RB(0, \sigma_\epsilon^2)$.

Com isso, o teste de Dickey-Fuller emprega as seguintes hipóteses:

$$H_0: \phi = 1;$$

$$H_1: \phi < 1.$$

E para o teste de hipóteses, a estatística $\hat{\tau}$ é utilizada:

$$\hat{\tau} = \frac{\hat{\phi} - 1}{S(\sum_{t=2}^n Z_{t-1}^2)^{-1/2}}, \quad (2.4)$$

onde $\hat{\phi}$ é o estimador de mínimos quadrados de ϕ (obtido a partir da regressão de Z_t sobre Z_{t-1}), n é o número de observações do processo e S^2 é determinado pela equação (2.5):

$$S^2 = \frac{\sum_{t=2}^n Z_t^2 - \hat{\phi} \sum_{t=2}^n Z_{t-1} Z_t}{n-2}. \quad (2.5)$$

A estatística $\hat{\tau}$ é comparada com valores críticos tabulados por Dickey e Fuller (1979) em função de n . Caso $\hat{\tau}$ seja menor do que o valor crítico, temos que a hipótese nula é rejeitada, e podemos assumir estacionariedade do processo.

O teste de Dickey-Fuller, entretanto, possui a limitação de considerar que o processo Z_t analisado é dependente apenas da observação imediatamente anterior a ele, ou Z_{t-1} . Dessa maneira, os mesmos autores propõem um método mais robusto para determinação da estacionariedade de uma série, chamado de Dickey-Fuller aumentado (ou teste ADF, devido à sua sigla em inglês). Este teste assume que o processo Z_t possui dependência também de termos anteriores a Z_{t-1} , sendo:

$$Z_t = \sum_{j=1}^{p+1} \varphi_j Z_{t-j} + a_t. \quad (2.6)$$

De acordo com Box, Jenkins, Reinsel e Ljung (2016), a estatística calculada passa a ser assintoticamente equivalente a:

$$\hat{\tau} = \frac{\hat{\rho} - 1}{S(\sum_{t=p+2}^n Z_{t-1}^2)^{-1/2}}, \quad (2.7)$$

onde $\hat{\rho} = \sum_{j=1}^{p+1} \hat{\varphi}_j$ e $\hat{\varphi}_j$ é o estimador de mínimos quadrados de φ_j em (2.6). Da mesma maneira, a hipótese nula deste teste é de que o processo em questão não é estacionário, e a estatística é comparada com valores críticos para $\hat{\tau}$ tabulados pelos mesmos autores. A estacionariedade do processo é aceita quando $\hat{\tau}$ assume valores inferiores aos dos valores críticos.

2.4. Normalidade

Uma outra suposição importante para o estudo e modelagem de séries temporais é a definição da distribuição de probabilidade da série e dos resíduos do modelo. Uma das principais suposições ao realizar a análise de séries temporais é de que o processo estudado adere à distribuição normal (ou gaussiana).

A variável aleatória X tem distribuição normal se sua função densidade de probabilidade é dada por:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty, \quad (2.8)$$

onde μ é a média de X , dada por

$$\mu = E\{X\},$$

e σ^2 é a variância de X , dada por

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

Assim, temos a notação $X \sim N(\mu, \sigma^2)$ para a variável aleatória X aderente à distribuição normal com parâmetros μ e σ^2 .

Aplicando este conceito de normalidade a processos estocásticos, Morettin e Tolo (2006) estabelecem a seguinte definição:

Definição: um processo estocástico $Z = \{Z(t), t \in T\}$ diz-se gaussiano se, para qualquer conjunto $t_1, t_2, \dots, t_n$ de T , as variáveis aleatórias $Z(t_1), Z(t_2), \dots, Z(t_n)$ têm distribuição normal n-variada.

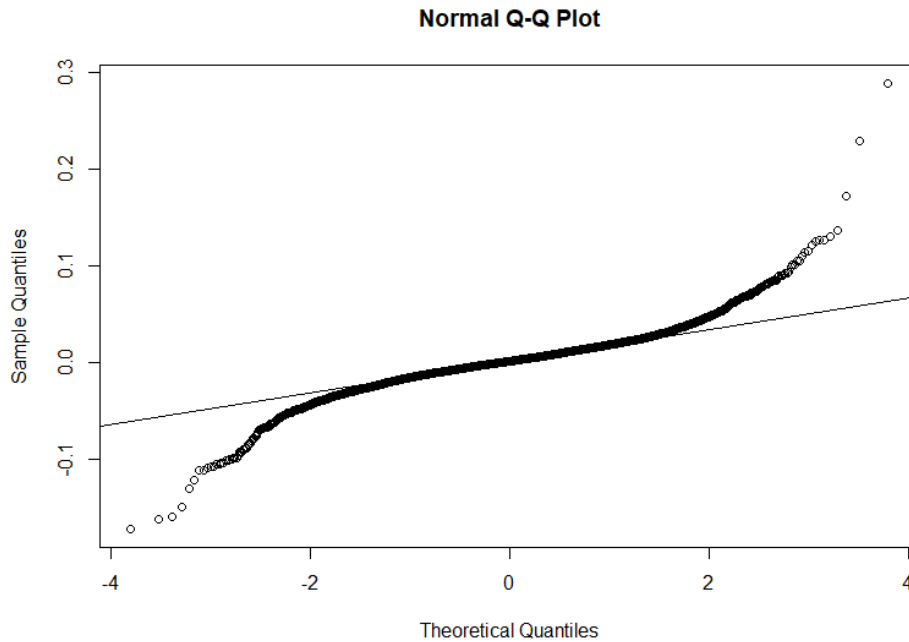
Um processo gaussiano é determinado pelas suas médias e covariâncias, propriedades que serão estudadas ao realizar a análise de séries temporais. Outra propriedade importante é que processos fracamente estacionários aderentes à distribuição normal são também estritamente estacionários, tendo, portanto, variância constante para todo $t \in T$.

No presente trabalho, serão utilizadas duas ferramentas para testar a normalidade da série a ser modelada e de seus resíduos: o gráfico quantil-quantil (ou apenas QQ) e o teste estatístico de Shapiro-Wilk.

A primeira delas é uma ferramenta visual, que permite observar se a série em questão se aproxima da distribuição normal graficamente. No gráfico QQ, são colocados em um gráfico de dispersão os quantis empíricos da amostra no eixo y, contra os quantis teóricos da distribuição normal padrão no eixo x. Quanto mais os dados da série aproximarem-se da reta, mais aderente é o processo a essa distribuição. Dessa maneira, se o processo testado for gerado a partir de uma distribuição normal, em teoria todas as suas observações estariam exatamente sobre a reta.

Na Figura 3, está ilustrado o gráfico QQ dos retornos diários do índice da Bolsa de Valores de São Paulo (Ibovespa). Visualmente, a série não parece ser aderente à distribuição normal, dado que diversos de seus pontos estão muito distantes da reta.

Figura 3 - Gráfico QQ da série de retornos diários do Ibovespa



Fonte: Elaborado pelo autor.

A segunda ferramenta utilizada é um teste estatístico, desenvolvido por Shapiro e Wilk (1965), que busca testar se uma série $Z_t = \{Z_1, Z_2, \dots, Z_n\}$ é aderente à distribuição normal. Dessa maneira, as hipóteses a serem do teste são as seguintes:

H_0 : os dados são aderentes à distribuição normal;

H_1 : os dados não são aderentes à distribuição normal.

Para tal, é calculada a estatística W , dada por:

$$W = \frac{\sum_{i=1}^n (a_i Z_i)^2}{\sum_{i=1}^n (Z_i - \bar{Z})^2}, \quad (2.9)$$

onde os termos a_i são determinados a partir da tabela de coeficientes de Shapiro-Wilk (ver Shapiro e Wilk (1965)) e $\bar{Z}$ é a média das observações do processo. Para verificar se a hipótese nula deve ser rejeitada, a estatística W é comparada com os valores críticos de $W_{n;\alpha}$, tabelados por Shapiro e Wilk, onde α é o nível de significância desejado.

De acordo com Yap e Sim (2011), o teste de Shapiro-Wilk é indicado para amostras não muito grandes, de até 50 observações, mas costuma ter resultados satisfatórios para séries com até 2000 dados.

2.5. Independência

O conceito de independência aplicado a processos estocásticos está relacionado à relação de uma observação do processo com suas observações antecedentes no

tempo. No estudo de séries temporais, três funções são essenciais para a determinação da independência das observações: as funções de autocovariância, autocorrelação e autocorrelação parcial. A partir da análise destas, é possível determinar se existe correlação entre os valores da série temporal espaçados de uma distância específica no tempo.

Nesta seção serão introduzidos os conceitos de função autocorrelação (fac) e função autocorrelação parcial (facp), e será detalhada a facv, já introduzida anteriormente na Seção 2.2.

2.5.1. Função de Autocovariância

Para um processo estacionário $\{X_t, t \in Z\}$, com média $\mu = 0$ e facv $\gamma_t = E\{X_t X_{t+\tau}\}$, temos que a função de autocovariância γ_t satisfaz as seguintes propriedades:

- i. $\gamma_0 > 0$;
- ii. $\gamma_{-\tau} = \gamma_\tau$;
- iii. $|\gamma_\tau| \leq \gamma_0$;
- iv. γ_τ é não negativa definida, no sentido que:

$$\sum_{j=1}^n \sum_{k=1}^n a_j a_k \gamma_{\tau_j - \tau_k} \geq 0,$$

para quaisquer números reais $a_1, \dots, a_n$ e $\tau_1, \dots, \tau_n$ de Z .

A facv de um processo estocástico pode ser estimada pela facv amostral do processo, dada por:

$$\hat{\gamma}_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_{t+h} - \bar{X})(X_t - \bar{X}), \text{ para } h = 0, 1, \dots, n-1, \quad (2.10)$$

onde $\bar{X}$ é a média amostral do processo X_t .

2.5.2. Função de Autocorrelação

A fac é extremamente importante nos processos de análise e modelagem de séries temporais, sendo utilizada tanto na fase de identificação quanto na fase de diagnóstico do modelo aplicado à série. A fac de um processo estacionário X_t mede a correlação entre X_t e $X_{t+\tau}$, sendo definida por:

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}, \tau \in Z. \quad (2.11)$$

A fac ρ_τ tem as mesmas propriedades da facv γ_τ , exceto para ρ_0 , para o qual assume-se o valor 1.

Podemos obter a estimativa da fac, $\hat{r}_h$, a partir das estimativas $\hat{\gamma}_h$, como apresentado na equação (2.12):

$$\hat{r}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}, \text{ para } h = 0, 1, n-1. \quad (2.12)$$

2.5.3. Função de Autocorrelação parcial

Além da fac, Box, Jenkins e Reinsel (1994) propõem um instrumento adicional para auxiliar na identificação e diagnóstico de modelos para séries temporais: a facp. Esta estatística, assim como a fac, é uma medida da correlação entre as observações X_t e $X_{t+\tau}$. Segundo Cowpertwait e Metcalfe (2009), a autocorrelação parcial no *lag* τ é a correlação resultante entre X_t e $X_{t-\tau}$, depois de remover o efeito de qualquer correlação devido aos termos de *lags* menores, causados por $X_{t-1}, X_{t-2}, \dots, X_{t+\tau-1}$.

Para definir esta função, Morettin e Toloi (2006) denotam por ϕ_{kj} o j -ésimo coeficiente de um modelo $AR(k)$ (que será detalhado na Seção 2.6), de modo que ϕ_{kk} seja seu último coeficiente. Dessa maneira, temos que:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \dots + \phi_{kk}\rho_{j-k}, j = 1, \dots, k. \quad (2.13)$$

A partir disto, obtemos as equações de Yule-Walker:

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{k-2} \\ \vdots & & & & \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \dots & 1 \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \dots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \dots \\ \rho_k \end{pmatrix} \quad (2.14)$$

Ao resolver a equação (2.14) para $k = 1, 2, \dots$, temos:

$$\begin{aligned} \phi_{11} &= \rho_1, \\ \phi_{22} &= \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}, \end{aligned}$$

e, generalizando:

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|}, \quad (2.15)$$

onde P_k é a matriz de autocorrelações e P_k^* é a matriz P_k com a última coluna substituída pelo vetor de autocorrelações. A variável ϕ_{kk} é denominada a autocorrelação parcial de ordem k do processo X_t , e mede a facp entre X_t e X_{t-k} .

Para estimar a facp, Morettin e Tolo (2006) propõem um método utilizando as estimativas da fac. Assim, é necessário, nas equações de Yule-Walker, substituir todas as fac ρ_j pelas suas estimativas r_j .

2.6. Modelos paramétricos

A classe de modelos que será estudada no presente trabalho e utilizada para modelar a série temporal do caso real é a de modelos paramétricos conhecida como abordagem de Box e Jenkins (1970). Nessa metodologia, os modelos possuem duas partes principais: a autorregressiva e a de média móvel, que são ajustadas ao conjunto de dados de forma a minimizar o erro quadrático médio entre a previsão e a observação. Além disso, serão estudados também os modelos sazonais para séries temporais, muito úteis na análise de séries mensais que costumam apresentar um padrão ao longo do ano.

Os modelos dessa classe são construídos a partir de um ciclo iterativo, tomando como base a estrutura dos dados a serem modelados. A partir de um primeiro passo, de especificação, na qual uma classe de modelos a ser utilizada é definida, os passos deste ciclo são realizados iterativamente na sequência:

- i. identificação;
- ii. estimação;
- iii. diagnóstico.

Estes passos serão descritos no Capítulo 4, da metodologia. O ciclo é repetido até que o modelo encontrado seja adequado aos dados utilizados e forneça um resultado satisfatório ao problema proposto. Outra estratégia que pode ser utilizada é realizar o estudo de diversos modelos através destes passos para compará-los posteriormente e decidir pelo melhor.

Segundo Morettin e Tolo (2006), os modelos devem ser o mais parcimonioso possível, ou seja, o número de parâmetros utilizados deve ser o menor possível que seja suficiente para uma boa modelagem. Dessa maneira, as previsões podem ser mais precisas, evitando problemas como sobreajuste (ou *overfitting*, do termo em inglês).

Há duas notações ainda não introduzidas que serão necessárias para a definição dos modelos paramétricos de Box e Jenkins (1970): o operador de translação para o passado, denotado por B , e o operador diferença, denotado por Δ .

O operador translação para o passado é definido por:

$$B^m Z_t = Z_{t-m}. \quad (2.16)$$

Enquanto o operador diferença é dado por:

$$\Delta Z_t = Z_t - Z_{t-1} = (1 - B)Z_t. \quad (2.17)$$

Para dados sazonais mensais, torna-se relevante o operador diferença de ordem 12, ou Δ^{12} :

$$\Delta^{12} Z_t = Z_t - Z_{t-12}.$$

2.6.1. Modelos autorregressivos (AR)

Seja Z_t uma série temporal com média μ . Os modelos autorregressivos buscam modelar a observação Z_t a partir das observações passadas da série $Z_{t-1}, Z_{t-2}, \dots, Z_{t-p}$, onde p é a ordem do modelo autorregressivo $AR(p)$. O processo $AR(p)$ pode ser definido por:

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \dots + \phi_p \tilde{Z}_{t-p} + a_t, \quad (2.18)$$

onde $\tilde{Z}_t = Z_t - \mu$. Os coeficientes ϕ_i , $i = 1, \dots, p$ são os pesos atribuídos pelo modelo às observações passadas, enquanto a_t é o ruído branco $RB(0, \sigma^2)$ no período t .

Morettin e Toloi (2006) ainda definem uma segunda forma para representar o modelo $AR(p)$, utilizando os coeficientes autorregressivos ϕ_i e o operador de translação para o passado B .

$$\phi(B) \tilde{Z}_t = a_t, \quad (2.19)$$

onde $\phi(B)$ é o operador autorregressivo de ordem p :

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p.$$

Ainda de acordo com Morettin e Toloi (2006), a fac de um processo autorregressivo decai seguindo uma mistura de polinômios, exponenciais e senoides amortecidas. Por outro lado, a facp de processos autorregressivos é significativa apenas para ordens maiores que p , de maneira que $\phi_k = 0$, para $k > p$.

2.6.2. Modelos de médias móveis (MA)

Os modelos de médias móveis, por outro lado, buscam modelar a série temporal Z_t a partir do ruído branco das observações passadas. Dessa maneira, um modelo de médias móveis de ordem q , denotado por $MA(q)$, utiliza como base para modelagem da observação Z_t os ruídos brancos $a_{t-1}, a_{t-2}, \dots, a_{t-q}$. Assim, o processo $MA(q)$ pode ser definido por:

$$Z_t = \mu + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}. \quad (2.20)$$

Usando uma notação semelhante à utilizada para o processo $AR(p)$, com o operador de translação para o passado B , temos:

$$\tilde{Z}_t = \theta(B)a_t, \quad (2.21)$$

onde

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

é o chamado operador de médias móveis de ordem q .

A fac de um modelo de médias móveis tem comportamento semelhante à facp de um processo autorregressivo. Em um modelo $MA(q)$, para ordens k maiores que q , temos que $\rho_k = 0$. Por outro lado, o comportamento da facp de processos de médias móveis tem um decaimento de seu valor absoluto de maneira exponencial, ou através de senoides amortecidas.

2.6.3. Modelos autorregressivos de médias móveis (ARMA)

Muitas séries encontradas na prática não possuem um comportamento totalmente aderente a um modelo específico autorregressivo ou de médias móveis de ordens baixas. De maneira a manter os modelos relativamente parcimoniosos, uma solução eficiente proposta por Box e Jenkins (1970) é a combinação de ambos os termos em um modelo autorregressivo de médias móveis, ou $ARMA(p, q)$. Os modelos desta classe são expressos como:

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \dots + \phi_p \tilde{Z}_{t-p} + a_t - \phi_1 a_{t-1} - \dots - \phi_q a_{t-q}. \quad (2.22)$$

Usando a notação com os operadores autorregressivo de ordem p , $\phi(B)$, e de médias móveis de ordem q , $\theta(B)$, temos:

$$\phi(B)\tilde{Z}_t = \theta(B)a_t. \quad (2.23)$$

Alguns casos interessantes de modelo $ARMA(p, q)$ são encontrados quando algum dos parâmetros do modelo é igual a 0. Dessa maneira, temos:

- i. $ARMA(p, 0) = AR(p)$;
- ii. $ARMA(0, q) = MA(q)$.

De acordo com Morettin e Toloi (2006), a fac de um processo $ARMA(p, q)$ é dada pela seguinte equação:

$$\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} + \dots + \phi_p \rho_{j-p}, j > q, \quad (2.24)$$

de maneira que é possível deduzir algumas informações a respeito do comportamento dessa propriedade. Para os *lags* menores do que q , portanto, a *fac* será afetada diretamente pelo parâmetro de médias móveis. Já para os *lags* maiores, o comportamento da *fac* é semelhante ao dos modelos autorregressivos.

Enquanto isso, a *facp* de processos $ARMA(p, q)$ tem um comportamento semelhante ao da *facp* de processos $MA(q)$ puros, ou seja, com decaimento determinado por exponenciais ou senoides amortecidas.

2.6.4. Modelos autorregressivos integrados de médias móveis (ARIMA)

Os modelos AR, MA e ARMA são apropriados para descrever séries temporais estacionárias, cujas médias são constantes ao longo do tempo. Porém, na prática, diversas séries não possuem essa propriedade, inclusive a maioria das séries econômicas e financeiras.

Para que os modelos estudados anteriormente sejam aplicáveis a séries não estacionárias, faz-se necessário o uso da diferenciação. Através dessa operação, é possível criar séries estacionárias a partir daquelas que não o são, utilizando o operador diferença. Tomando Z_t como uma série não estacionária, temos que

$$W_t = Z_t - Z_{t-1} = (1 - B)Z_t = \Delta Z_t$$

pode ser estacionária. Esse procedimento de diferenciação pode ser realizado sucessivamente, até que a série:

$$W_t = \Delta^d Z_t$$

seja estacionária. De acordo com Morettin e Toloi (2006), na maior parte dos casos basta a realização de apenas uma ou duas diferenciações para que a série obtida seja estacionária.

A partir disso, surgem os modelos ARIMA, que correspondem à aplicação do modelo ARMA à série estacionária obtida a partir da série não estacionária. Aplicando o modelo $ARMA(p, q)$ à série estacionária W_t , obtemos a seguinte equação:

$$\phi(B)W_t = \theta(B)a_t. \quad (2.25)$$

De maneira que

$$\phi(B)\Delta^d Z_t = \theta(B)a_t \quad (2.26)$$

é um modelo $ARIMA(p, d, q)$ de Z_t . De maneira semelhante aos modelos $ARMA(p, q)$, os coeficientes p e q são, respectivamente, as ordens de $\phi(B)$ e $\theta(B)$, enquanto o coeficiente d corresponde à ordem de diferenciação da série não estacionária.

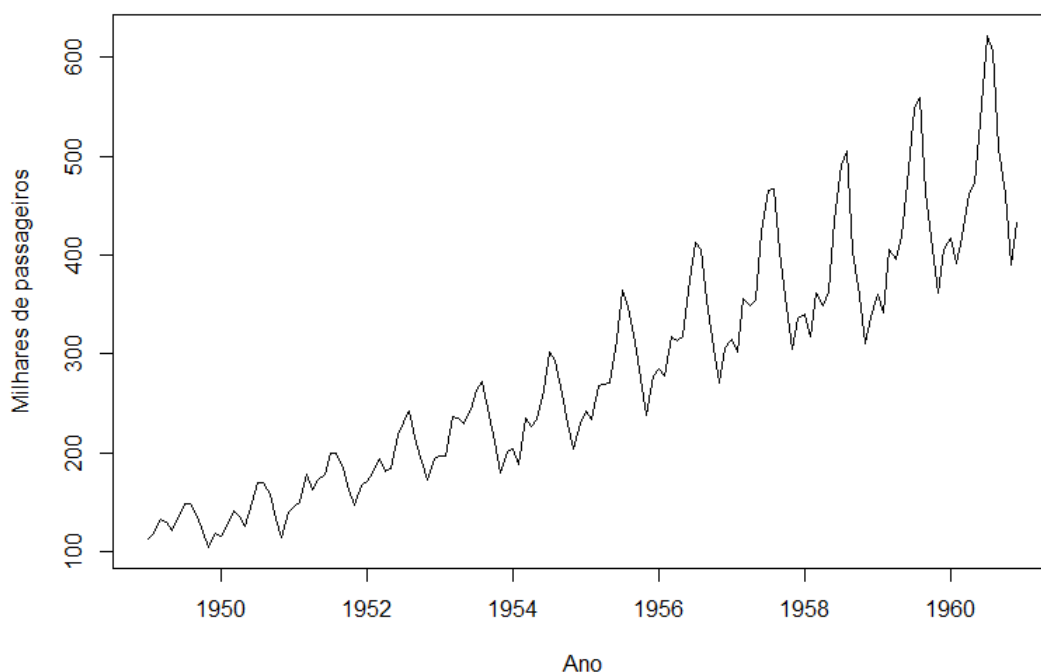
Após a etapa de diferenciação, o comportamento de modelos $ARIMA(p, d, q)$ é semelhante ao de modelos $ARMA(p, q)$. Com isso, fica evidente que:

$$ARIMA(p, 0, q) = ARMA(p, q).$$

2.6.5. Modelos ARIMA sazonais (SARIMA)

Na prática, diversas séries temporais, especialmente as de periodicidade mensal, podem apresentar comportamento sazonal. Em séries com esta característica, ocorrem picos ou vales nos valores das observações em determinadas épocas do ano, causadas por fatores externos à série. Box, Jenkins e Reinsel (1994) trazem como exemplo de sazonalidade a série do número de passageiros aéreos internacionais, de 1949 a 1960, mostrada na Figura 4.

Figura 4 - Série temporal de passageiros internacionais no transporte aéreo



Fonte: Elaborado pelo autor.

Como podemos notar, há claramente uma periodicidade nos valores da série, que tendem a ser mais altos nos meses de junho, julho e agosto, períodos que coincidem com as férias escolares na maioria dos países do hemisfério norte.

De acordo com Morettin e Toloi (2006), quando o processo Z_t apresenta comportamento sazonal, um modelo que pode ser útil é o seguinte:

$$Z_t = \mu_t + N_t, \quad (2.27)$$

onde N_t é um processo estacionário que pode ser modelado por um ARMA(p, q) e μ_t é um processo estocástico periódico, que satisfaz

$$(1 - B^{12})\mu_t = Y_t. \quad (2.28)$$

Na equação (2.28), Y_t também é um processo estacionário da forma ARMA(p, q).

Se aplicarmos o operador $(1 - B^{12})$ à equação (2.27), temos

$$(1 - B^{12})Z_t = (1 - B^{12})\mu_t + (1 - B^{12})N_t, \quad (2.29)$$

que pode ser simplificado para

$$(1 - B^{12})Z_t = Y_t + (1 - B^{12})N_t. \quad (2.30)$$

Relembrando que ambos N_t e Y_t são processos ARMA(p, q), temos:

$$\phi_Y(B)Y_t = \theta_Y(B)a_t, \quad (2.31)$$

$$\phi_N(B)N_t = \theta_N(B)\epsilon_t, \quad (2.32)$$

onde a_t e ϵ_t são ruídos brancos independentes.

Assim, pode-se demonstrar que a equação (2.30) equivale a:

$$(1 - \Phi_1 B^{12} - \dots - \Phi_P B^{12P})(1 - B^{12})^D Z_t = (1 - \Theta_1 B^{12} - \dots - \Theta_Q B^{12Q})\alpha_t. \quad (2.33)$$

Ou, simplificadamente,

$$\Phi(B^{12})\Delta_{12}^D Z_t = \Theta(B^{12})\alpha_t, \quad (2.34)$$

onde

- $\Phi(B^{12}) = 1 - \Phi_1 B^{12} - \dots - \Phi_P B^{12P}$ é o operador autorregressivo sazonal de ordem P , estacionário;
- $\Theta(B^{12}) = 1 - \Theta_1 B^{12} - \dots - \Theta_Q B^{12Q}$ é o operador de médias móveis sazonal de ordem Q , invertível;
- $\Delta_{12} = (1 - B^{12})$ é o operador de diferença sazonal;
- D indica o número de diferenças sazonais aplicadas;
- e α_t satisfaz um modelo ARIMA(p, d, q) ou é ruído branco.

No último caso, a fac do processo Z_t é 0 para todos os lags não sazonais (isto é, que não são múltiplos de P) e o modelo determinado pela equação (2.34) é denominado um modelo sazonal puro. Caso α_t satisfaça um modelo ARIMA(p, d, q), temos, conforme determinado na seção 2.6.4:

$$\phi(B)\Delta^d \alpha_t = \theta(B)a_t. \quad (2.35)$$

Com isso, podemos demonstrar que

$$\phi(B)\Phi(B^{12})\Delta^d \Delta_{12}^D Z_t = \theta(B)\Theta(B^{12})a_t, \quad (2.36)$$

onde

- $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ é o operador autorregressivo de ordem p ;
- $\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ é o operador de médias móveis de ordem q ;
- d é o número de diferenças de ordem 1 aplicadas à Z_t ;

com $\Phi(B^{12})$, $\Theta(B^{12})$, D e a_t já tendo sido definidos anteriormente.

O modelo determinado pela equação (2.36) é denominado ARIMA sazonal multiplicativo, ou apenas SARIMA, de ordem $(p, d, q) \times (P, D, Q)_{12}$.

3. REVISÃO BIBLIOGRÁFICA SOBRE CONTROLE ESTATÍSTICO DE PROCESSO

O Capítulo 3 tem como objetivo apresentar e aprofundar-se na base teórica do CEP, e mais especificamente nos gráficos de controle, uma das ferramentas que o compõem e que será utilizada para monitorar um processo real no Capítulo 5, a partir de sua modelagem por séries temporais. Ao longo deste capítulo, portanto, serão introduzidos os principais conceitos do CEP e será explicitado o funcionamento dos principais tipos de gráficos de controle, além de como eles se relacionam com processos modelados por séries temporais.

3.1. Introdução ao CEP

Para que um produto atenda às expectativas de seus clientes, é essencial que ele seja produzido por um processo estável e repetível, de maneira que ele tenha características aderentes às que foram definidas no seu projeto, com pouca variabilidade sobre as dimensões alvo e a qualidade.

De acordo com Montgomery (2012), o CEP surge como uma poderosa coleção de ferramentas para a solução de problemas relacionados à estabilidade de processos e melhora da capacidade através da redução da variabilidade. O mesmo autor define o CEP como um dos principais avanços tecnológicos do século XX, por ser baseado em princípios sólidos, ser fácil de usar e poder ser aplicado a qualquer processo. As sete principais ferramentas que compõem o CEP são:

1. Histograma;
2. Lista de verificação;
3. Gráfico de Pareto;
4. Diagrama de causa e efeito;
5. Diagrama de concentração de defeito;
6. Diagrama de dispersão;
7. Gráfico de controle.

Destas ferramentas, o gráfico de controle destaca-se como a mais sofisticada, e será a utilizada para analisar a variabilidade da série do IBC-Br no presente trabalho. Assim, nas próximas seções, estudaremos mais a fundo os gráficos de controle, suas características e seus principais tipos.

Para compreender os conceitos estatísticos que formam a base do CEP, entretanto, ainda é necessário descrever a teoria da variabilidade de Shewhart, desenvolvida por Walter Shewhart ao longo da década de 1920. De acordo com ela, qualquer processo produtivo, independentemente de quão bem projetado e mantido ele seja, terá uma variabilidade natural inerente a ele, que não pode ser eliminada totalmente, resultado de múltiplas causas pequenas e inevitáveis. Quando o processo está operando apenas com estas causas naturais de variação, dizemos que ele está sob controle estatístico.

Entretanto, outros tipos de variabilidade também podem estar presentes, ocasionalmente, na saída de um processo, que costumam ter como origem três motivos: máquinas mal ajustadas ou controladas, erros de operação ou insumos defeituosos. Estas são chamadas de causas especiais, e quando um processo está operando na presença delas, dizemos que ele está fora de controle.

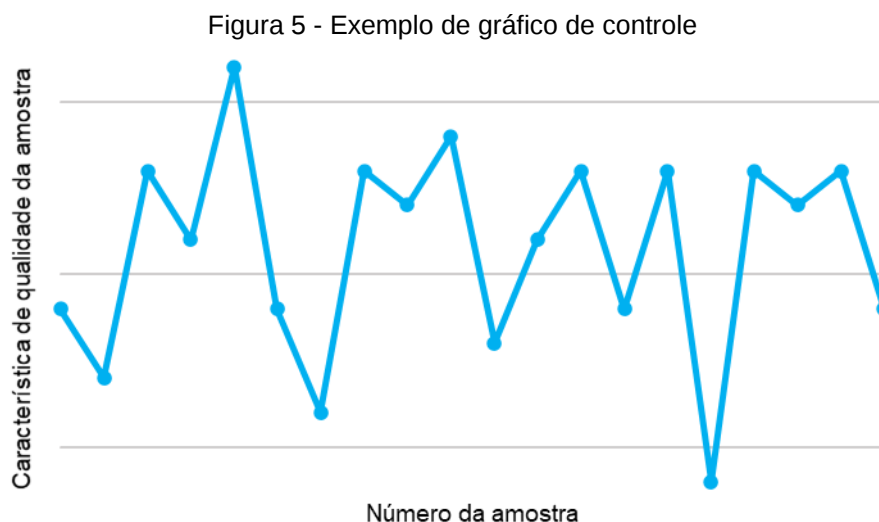
3.2. Gráficos de controle

No contexto da variabilidade dos processos, os gráficos de controle têm como finalidade a rápida detecção de causas especiais através da verificação de se o processo está ou não sob controle, de modo que sejam tomadas ações sobre elas o mais cedo possível e que menos peças sejam afetadas. Com isso, é possível atingir o principal objetivo do CEP, que é a eliminação da variabilidade do processo. Como já mencionado, pode não ser possível eliminá-la completamente, mas espera-se que ao menos as causas especiais possam ser.

De acordo com Montgomery (2012), gráfico de controle é uma disposição gráfica de uma característica de qualidade que foi medida ou computada a partir de uma amostra contra o número da amostra ou o tempo. O gráfico contém uma linha central (LC) que representa o valor médio da característica de qualidade quando o processo está sob controle. Outras duas linhas horizontais, chamadas de limite superior de controle (LSC) e limite inferior de controle (LIC), também são mostradas no gráfico. Estes limites de controle são escolhidos de maneira que, se o processo está sob controle, quase todas as amostras vão ficar entre eles.

Assim, enquanto os pontos estão dentro destes limites, o processo é assumido como sob controle, e não é necessária nenhuma tomada de ação. No momento em que uma amostra fica acima do LSC ou abaixo do LIC, assume-se que o processo está fora de controle, e passa a ser necessário realizar ações para a identificação e

correção da causa especial que está atuando sobre ele. Um exemplo de gráfico de controle pode ser visualizado na Figura 5.



Fonte: Elaborado pelo autor.

Montgomery (2012) afirma que há uma relação muito próxima entre gráficos de controle e testes de hipóteses. Podemos visualizar os gráficos de controle como testes de hipótese aplicados múltiplas vezes ao processo, com cada amostra sendo avaliada pelos limites de controle a partir das seguintes hipóteses:

H_0 : O processo monitorado está sob controle;

H_1 : O processo monitorado não está sob controle.

Nesse caso, a hipótese nula é rejeitada se a amostra se encontrar fora do intervalo entre os limites de controle. Caso contrário, a hipótese nula é aceita, e o processo é considerado sob controle.

A visão do gráfico de controle como um teste de hipóteses contínuo é muito útil ao se avaliar o desempenho da ferramenta. Para tal, é importante relembrar os dois tipos de erro que podem ocorrer ao longo do processo de tomada de decisão determinado por um teste de hipótese: erro do tipo I e erro do tipo II.

No primeiro, a hipótese nula é rejeitada mesmo sendo verdadeira. No caso dos gráficos de controle, isso significa que o processo analisado está sob controle, mas os limites de controle apontam que ele não está. Assim, a probabilidade de que este erro (também denominado “alarme falso”) aconteça ao se monitorar um processo Z_t , $t \in N$, é de:

$$\alpha = P[\text{erro tipo I}] = P[Z_t > LSC | \text{processo sob controle}] + P[Z_t < LIC | \text{processo sob controle}]$$

Enquanto isso, o erro do tipo II é caracterizado quando o teste de hipótese aceita a hipótese nula com ela sendo falsa. No caso do gráfico de controle, portanto, ele acontece quando o processo analisado se encontra dentro dos limites de controle determinados, mas não está sob controle. A probabilidade de ocorrência do “erro de não-detecção” é dada por:

$$\beta = P[\text{erro tipo II}] = P[LIC < Z_t < LSC | \text{processo fora de controle}]$$

Evidentemente, a probabilidade de ocorrência dos erros do tipo I e II são diretamente relacionadas à amplitude dos limites de controle determinados. Com o aumento da amplitude, a tendência é que α seja menor e β maior, e vice-versa. De acordo com o parâmetro de desempenho desejado para cada uma das probabilidades, os limites devem ser calculados. Montgomery (2012) estabelece um modelo geral para gráficos de controles, dada a estatística amostral w (que mede uma característica de interesse), sua média μ_w e seu desvio padrão σ_w :

$$LSC = \mu_w + L\sigma_w; \quad (3.1)$$

$$LC = \mu_w; \quad (3.2)$$

$$LIC = \mu_w - L\sigma_w. \quad (3.3)$$

Dessa maneira, o parâmetro L define a distância dos limites de controle para a LC, e a amplitude ao longo da qual o processo é considerado sob controle.

Assim, resta apenas compreender como o valor de L é definido a partir dos valores desejados para α e β . Uma métrica comumente utilizada em gráficos de controle é o ARL (da sigla em inglês para “comprimento médio da sequência”). Montgomery (2012) define o ARL como a quantidade média de pontos que precisam ser plotados no gráfico antes que um deles indique um processo fora de controle. O cálculo do ARL é definido pela seguinte equação:

$$ARL = \frac{1}{p}, \quad (3.3)$$

onde p é a probabilidade de que um ponto exceda os limites de controle.

Há duas estatísticas importantes relacionadas a este conceito. A primeira delas é o ARL_0 , que indica o comprimento médio da sequência quando o processo está sob controle, e é dada por:

$$ARL_0 = \frac{1}{\alpha}. \quad (3.4)$$

A segunda delas é o ARL_1 , que indica a demora para se identificar um processo fora de controle, sendo definida por:

$$ARL_1 = \frac{1}{1 - \beta}. \quad (3.5)$$

O ARL pode ser definido analiticamente quando a estatística monitorada é conhecida. Caso contrário, é necessário realizar simulações da série em questão para se obter os valores de ARL_0 e ARL_1 .

3.3. Tipos de gráficos de controle

Existem diversos gráficos controles distintos, com cada um tendo sua própria metodologia de monitoramento dos processos e possuindo características que os tornam mais adequados para determinadas situações. Há dois tipos principais de gráficos de controles: para variáveis e para atributos. O primeiro monitora características de qualidade mensuráveis numericamente em uma escala contínua, como o comprimento e o volume, enquanto o segundo costuma monitorar características não passíveis de medição numérica, como itens defeituosos e não defeituosos. No presente trabalho, nos interessam apenas os gráficos de controle para variáveis. Dentre eles, destacam-se os três a seguir, que serão aprofundados ao longo desta seção:

- i. Gráficos de Shewhart;
- ii. Gráfico de somas acumuladas (CUSUM, da sigla em inglês);
- iii. Gráfico de média móvel exponencialmente ponderada (EWMA, da sigla em inglês).

Os gráficos de controle de Shewhart utilizam como critério de decisão para determinar se o processo está sob ou fora de controle apenas a estatística obtida com a última amostra. Esta característica é citada por Montgomery (2012) como uma grande desvantagem deste tipo de gráfico, dado que ele ignora qualquer informação dada pela sequência anterior das estatísticas. Por conta disso, o gráfico de Shewhart acaba sendo menos sensível a pequenas mudanças no parâmetro da variável de controle mensurada. Por outro lado, este tipo de gráfico acaba sendo um dos mais utilizados para o monitoramento de processos estatísticos, por conta de sua fácil implementação e manutenção.

Os gráficos CUSUM e EWMA destacam-se como duas alternativas eficientes para os de Shewhart, sendo caracterizados como gráficos de controle com memória, dado que levam em consideração informações dadas pelas observações anteriores, não apenas pela mais recente. Montgomery (2012) afirma que estes dois tipos de gráficos de controle costumam ser mais eficientes para detectar pequenos desvios na média do processo, o que pode torná-los mais adequados para o monitoramento de séries econômicas e financeiras.

3.3.1. Shewhart

Os gráficos de Shewhart podem ser classificados de acordo com a estatística que é controlada por ele, das quais se destacam a média e a amplitude amostrais. Assim, apresentaremos a seguir estes dois gráficos.

Para tal, Montgomery (2012) assume uma característica de qualidade com distribuição normal, média μ e desvio padrão σ , com μ e σ ambos conhecidos. Dada uma amostra de tamanho n , temos:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad (3.6)$$

com $\bar{x}$ também sendo distribuído normalmente e tendo média μ e desvio padrão $\sigma_x = \sigma/\sqrt{n}$. Assim, os limites de controle são dados por:

$$LSC = \mu + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \quad (3.7)$$

$$LC = \mu; \quad (3.8)$$

$$LIC = \mu - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad (3.9)$$

onde α é a probabilidade de a média amostral cair fora do intervalo determinado pelos limites.

Entretanto, na prática, é comum que os parâmetros μ e σ não sejam conhecidos. Nesse caso, é necessário utilizar estimadores para a construção do gráfico de Shewhart. Supondo que m amostras estejam disponíveis, cada uma com n observações da característica mensurada, podemos estimar μ como a média das médias amostrais, dada por:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m}. \quad (3.10)$$

Resta, ainda, uma estimativa do desvio padrão σ . De acordo com Montgomery (2012), ela pode ser obtida a partir da amplitude das m amostras. Para cada amostra, temos que a amplitude é dada por:

$$R = x_{max} - x_{min}. \quad (3.11)$$

A amplitude média, então, é dada por:

$$\bar{R} = \frac{R_1 + R_2 + \dots + R_m}{m}. \quad (3.12)$$

Com isso, os limites de controle do gráfico para a média, também chamado de gráfico $\bar{x}$, passam a ser determinados por:

$$LSC = \bar{\bar{x}} + A_2 \bar{R}; \quad (3.13)$$

$$LC = \bar{\bar{x}}; \quad (3.14)$$

$$LIC = \bar{\bar{x}} - A_2 \bar{R}, \quad (3.15)$$

com A_2 sendo uma constante em função de n tabulada por Montgomery (2012).

E os limites de controle do gráfico para a amplitude, ou gráfico R , são determinados por:

$$LSC = D_4 \bar{R}; \quad (3.16)$$

$$LC = \bar{R}; \quad (3.17)$$

$$LIC = D_3 \bar{R}, \quad (3.18)$$

onde D_4 e D_3 também são constantes em função de n tabuladas pelo mesmo autor.

Cabe ressaltar que os procedimentos apresentados por Montgomery (2012) são aplicáveis quando se conhece a distribuição da variável monitorada. No caso do presente trabalho, serão utilizadas simulações para a determinação destes limites de controle.

3.3.2. CUSUM

O gráfico de controle CUSUM tem como estatística de monitoramento a soma cumulativa de todos os desvios das observações de uma amostra com relação a um valor alvo. Dessa maneira, as informações relativas às amostras passadas são incorporadas ao cálculo da estatística, conforme sua natureza de ser um gráfico de controle com memória define.

Montgomery (2012), para definir o gráfico CUSUM, considera um processo x , cuja i -ésima observação é dada por x_i . Quando o processo está sob controle, x_i possui distribuição normal, com média μ_0 e desvio padrão σ . Assume-se que este último parâmetro é conhecido ou possui uma estimativa confiável. O valor μ_0 é definido

como valor alvo do processo, ou seja, o valor por volta do qual espera-se que o processo varie. Assim, o gráfico CUSUM assinala que o processo está fora de controle quando ele se distancia muito deste valor.

São definidas então duas estatísticas, uma para monitorar o limite superior do processo e outra para o inferior: C^+ e C^- , respectivamente. O valor C^+ , chamado de CUSUM unilateral superior, acumula as variações acima de μ_0 , enquanto C^- , ou CUSUM unilateral inferior, acumula as variações do processo abaixo de μ_0 . Dessa maneira, eles são calculados conforme as equações a seguir:

$$C_i^+ = \max[0, x_i - (\mu_0 + K) + C_{i-1}^+]; \quad (3.19)$$

$$C_i^- = \max[0, (\mu_0 - K) - x_i + C_{i-1}^-], \quad (3.20)$$

onde os valores iniciais são $C_0^+ = C_0^- = 0$. K é usualmente denominado valor de referência, e é calculado como o valor médio entre o alvo μ_0 e a média dos valores fora de controle μ_1 , que estamos interessados em detectar rapidamente. Assim:

$$K = \frac{|\mu_1 - \mu_0|}{2}. \quad (3.21)$$

Portanto, nota-se que C_i^+ e C_i^- acumulam os desvios do valor alvo μ_0 que forem maiores que K , e ambas as estatísticas são zeradas ao se tornarem negativas. Se qualquer uma delas exceder o intervalo de decisão H , o processo passa a ser considerado como fora de controle

Assim, a seleção deste parâmetro H é outra etapa importante para a construção do gráfico de controle CUSUM. Segundo Montgomery (2012), um valor razoável para H é de cinco vezes o desvio padrão σ . Outra maneira de definir este valor é através de simulações, a partir de um valor alvo do ARL definido para o gráfico.

3.3.3. EWMA

O gráfico de controle EWMA foi originalmente introduzido por Roberts (1959), sendo uma segunda alternativa para o gráfico de Shewhart quando se deseja detectar pequenas variações no processo. O EWMA costuma pode ser utilizado com observações individuais e apresenta desempenho semelhante ao do gráfico CUSUM.

Considerando novamente um processo x , cuja i -ésima observação é dada por x_i , temos que a estatística a ser monitorada é dada por:

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1}, \quad (3.22)$$

onde $0 < \lambda \leq 1$ é uma constante e o valor inicial de z_i é o próprio valor alvo do processo, ou seja:

$$z_0 = \mu_0.$$

A partir da equação (3.22), é possível notar que z_i é uma soma ponderada de todas as observações anteriores da amostra atual, de acordo com o valor definido por λ . No caso, o parâmetro λ define o peso dado à observação atual, enquanto as observações passadas são ponderadas pelo produto entre potências de λ e $(1 - \lambda)$. Portanto, esta constante deve ser definida de acordo com o peso que se deseja dar para a observação atual e para as anteriores. A equação (3.23) mostra como λ afeta a ponderação de cada observação:

$$\begin{aligned} z_i &= \lambda x_i + (1 - \lambda)[\lambda x_{i-1} + (1 - \lambda)z_{i-2}] \\ &= \lambda x_i + \lambda(1 - \lambda)x_{i-1} + (1 - \lambda)^2 z_{i-2}. \end{aligned}$$

Recursivamente, para z_{i-j} , $j = 2, 3, \dots, t$, temos:

$$z_i = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0. \quad (3.23)$$

De acordo com Montgomery (2012), quanto menor o valor de λ , mais fácil é de se detectar pequenas variações – entretanto, aumenta também a probabilidade de ocorrerem alarmes falsos.

O gráfico EWMA é construído a partir da plotagem dos valores z_i *versus* o número de amostra, ao contrário do Shewhart e CUSUM, nos quais o valor da própria variável é inserido no gráfico. Montgomery (2012) define os limites de controle para este gráfico como:

$$LSC = \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2 - \lambda)} [1 - (1 - \lambda)^{2i}]}; \quad (3.24)$$

$$LC = \mu_0; \quad (3.25)$$

$$LIC = \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2 - \lambda)} [1 - (1 - \lambda)^{2i}]}. \quad (3.26)$$

Nas equações (3.24), (3.25) e (3.26), σ é o desvio padrão da amostra e L é um fator que define a amplitude dos limites de controle. De acordo com Montgomery (2012), o valor de $L = 3$ costuma funcionar razoavelmente bem, especialmente para valores maiores de λ . Para λ menores, o intervalo entre 2,6 e 2,8 para L costuma apresentar bons resultados para o gráfico EWMA. Da mesma maneira que para a

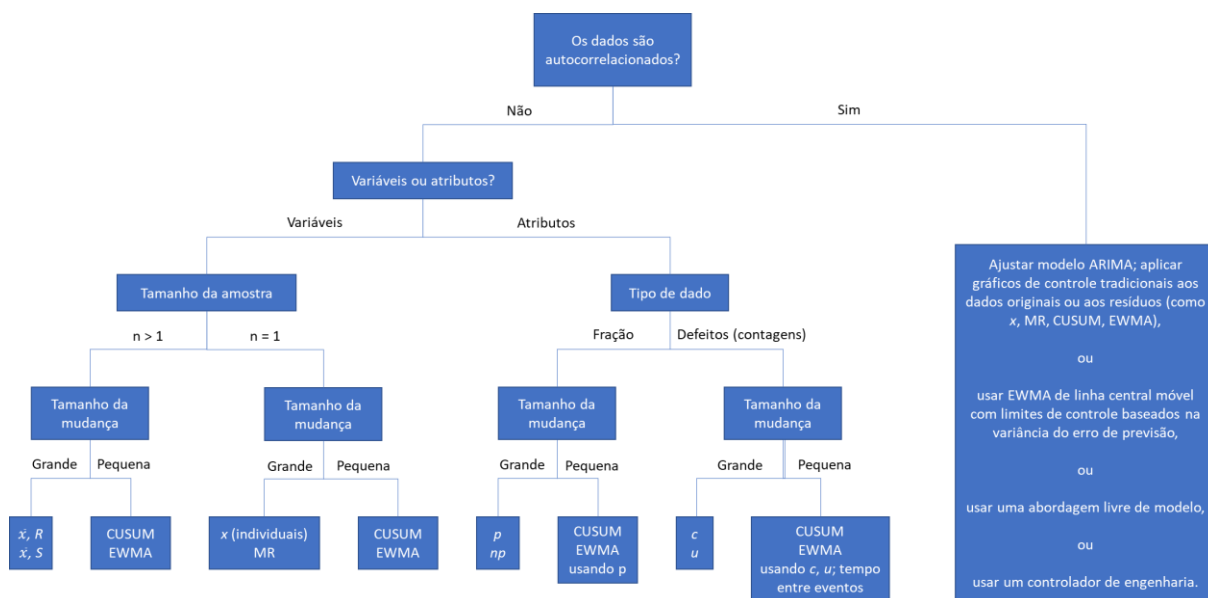
definição de H para o CUSUM, o valor de L pode ser determinado através de simulações do processo, visando atingir um valor alvo para o ARL.

3.4. Escolha do gráfico de controle

Existem vários tipos de gráficos de controle, conforme apresentado ao longo das Seções 3.2 e 3.3, com cada um tendo suas particularidades e sendo mais aplicável a determinado processo. Por essa razão, um passo muito importante ao se realizar o monitoramento de um processo estatístico é a determinação de qual será a categoria de gráfico de controle utilizada para esta finalidade.

Visando facilitar este processo de escolha, Montgomery (2012) propõe um fluxograma com diretrizes para a seleção do tipo de gráfico a ser utilizado, de acordo com a natureza da variável monitorada e das necessidades de tal controle. A Figura 6 apresenta o fluxograma de Montgomery.

Figura 6 - Fluxograma com diretrizes para a seleção do gráfico de controle



Fonte: Adaptado de Montgomery (2012).

Podemos notar que diversas variáveis relacionadas ao processo estudado serão utilizadas para este processo de tomada de decisão, como a autocorrelação entre os dados, o tipo de processo (variável ou atributo), o tamanho da amostra, entre outros.

Focando no caso de dados autocorrelacionados, o ramo da direita do fluxograma sugere quatro opções distintas:

- i. Ajustar modelo ARIMA e aplicar gráficos de controle tradicionais;

- ii. Usar um gráfico EWMA de linha central móvel;
- iii. Usar uma abordagem livre de modelo;
- iv. Usar um controlador de engenharia.

4. METODOLOGIA

O objetivo deste capítulo é explicitar de maneira detalhada a metodologia que será empregada para a construção de um modelo de previsão de séries temporais e para seu monitoramento através de gráficos de controle, baseada na revisão bibliográfica realizada nos dois capítulos anteriores. Ela será aplicada sobre uma série real, a do IBC-Br, no Capítulo 5. A metodologia a ser apresentada é aplicável a diversos casos reais, podendo ser replicada para o monitoramento de outras séries temporais com periodicidade anual.

4.1. Coleta e ajuste dos dados

O primeiro passo necessário para a aplicação da metodologia proposta é a coleta dos dados da série temporal que será utilizada na análise. É essencial que a fonte dos dados seja confiável, de maneira que o modelo seja ajustado sobre os dados corretos e seja aplicável ao caso real. Na maior parte dos casos, a fonte mais confiável que se pode obter é a da própria instituição que calcula e divulga os dados, como o IBGE para série do Índice Nacional de Preços ao Consumidor Amplo (IPCA) ou o Banco Central para o IBC-Br. Para séries econômicas e financeiras, outra fonte extremamente confiável é a *Bloomberg*, largamente utilizada em instituições que operam no mercado financeiro.

Outro ponto que deve ser ressaltado é que, para que o modelo seja eficiente, é importante que haja um número de observações suficiente para a estimação e modelagem dos dados com boa precisão. De acordo com Hanke e Wichern (2009), no caso de modelos sazonais com periodicidade mensal, é importante que se tenha ao menos de seis a 10 anos de observações, dependendo das especificações da série.

A partir da obtenção dos dados através de uma fonte confiável, a etapa seguinte é a análise deles para que estejam em um formato sobre o qual a modelagem SARIMA é aplicável. Dessa maneira, como demonstrado no Capítulo 2, a série a ser utilizada deve ser estacionária. Em séries econômicas, é comum que existam tendências, locais ou globais, e sazonalidade. Neste caso, é necessária a aplicação de transformações sobre a série, de forma que esses comportamentos sejam eliminados e que seja obtida uma série estacionária.

O presente trabalho propõe utilizar duas ferramentas para a verificação da estacionariedade da série temporal: a visualização gráfica da mesma e o teste ADF.

Na primeira, é necessário fazer a plotagem dos valores da série ao longo do tempo. A partir disso, é possível analisá-la visualmente de modo a verificar se existe algum tipo de não estacionariedade, como tendências ou sazonalidade.

O segundo deles é um teste estatístico, introduzido no Capítulo 2, que pode ser realizado através de programas e softwares de análise estatística, como o RStudio. Através dele, é possível testar a hipótese se a série em questão é estacionária ou não.

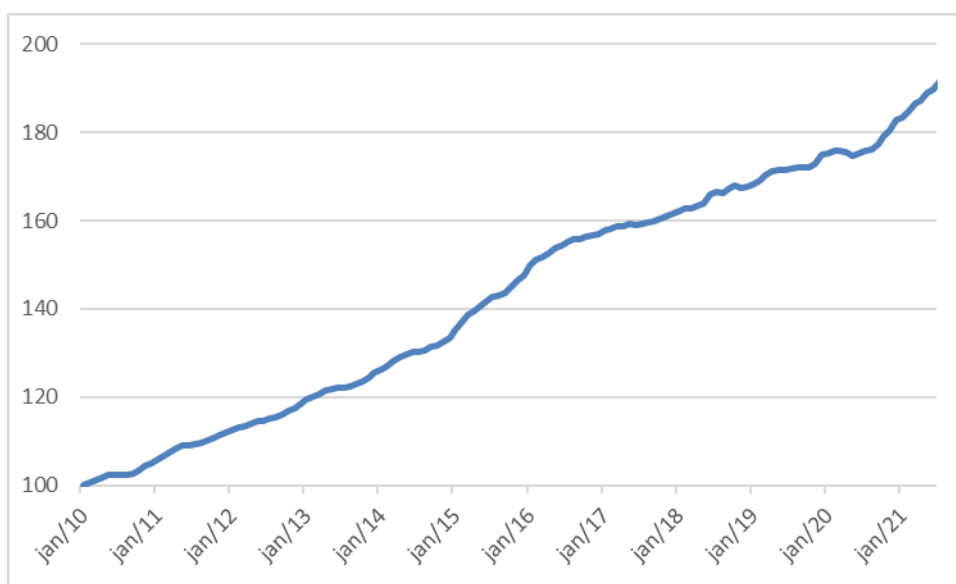
Caso a série analisada não seja estacionária, é necessário transformá-la até que ela o seja, para que seja possível aplicar a modelagem SARIMA sobre ela. Em séries econômicas e financeiras, é comum que seja utilizada a série de retornos ou de log-retornos dos dados, que costuma ser estacionária. Dada a série temporal X_t , $t \in Z$, para se obter a série de retornos R_t de X_t , a transformação (4.1) é realizada:

$$R_t = \frac{X_t}{X_{t-1}} - 1, \quad (4.1)$$

com os valores R_t geralmente expressos em porcentagem. No caso de séries mensais, os retornos são equivalentes à variação mensal das observações.

Na Figura 7, é possível visualizar graficamente a série do IPCA, de janeiro de 2010 a junho de 2021. Visualmente, fica claro que a série em questão não é estacionária, com uma tendência linear de aumento do índice desde seu início.

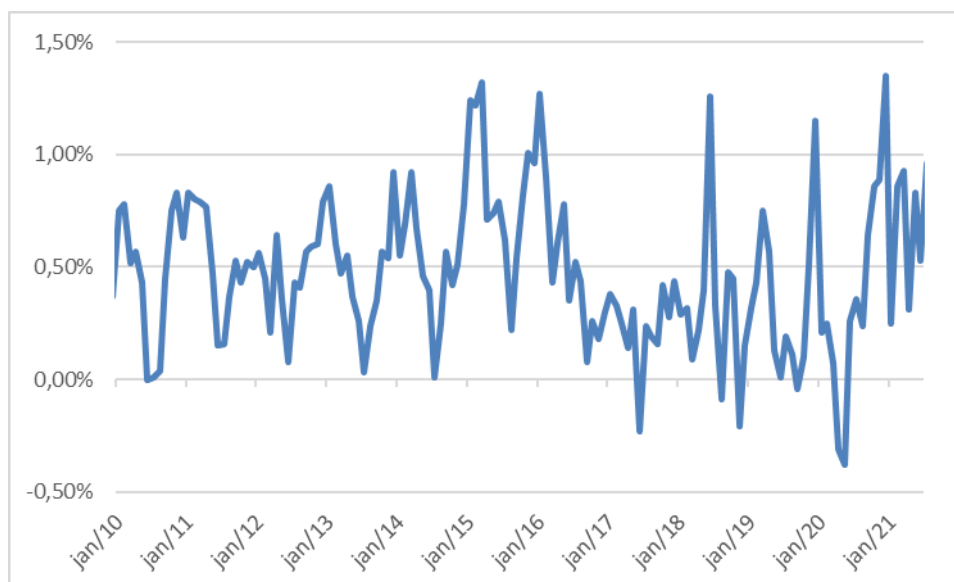
Figura 7 - Número-índice do IPCA



Fonte: Elaborado pelo autor.

Tomando os retornos dela, a série da Figura 8 é obtida. Como podemos notar, a tendência de crescimento foi eliminada, e a série parece variar em torno de uma média constante ao longo do tempo.

Figura 8 - Variação mensal do IPCA



Fonte: Elaborado pelo autor.

Com relação à sazonalidade, o método a ser utilizado para tratamento da série é o de diferenciação, já introduzido no Capítulo 2. A aplicação de apenas uma diferenciação sazonal na série de retornos costuma ser suficiente para que seja possível modelá-la utilizando modelos autorregressivos e de médias móveis. Dessa maneira, a série que será modelada é D_t , que no caso de periodicidade mensal é obtida por:

$$D_t = R_t - R_{t-12}. \quad (4.2)$$

A distribuição da série D_t é outro ponto que será analisado antes de sua modelagem, de forma a verificar se ela segue uma distribuição normal. Para tal, serão utilizados o gráfico QQ e o teste de Shapiro-Wilk, introduzidos na Seção 2.4, e um histograma da série.

A partir disso, com a série devidamente ajustada para a modelagem ARMA, será possível partir para os passos iterativos da abordagem Box e Jenkins, começando pela identificação do modelo.

4.2. Identificação do modelo

Como introduzido na Seção 2.6, a primeira etapa do ciclo iterativo da abordagem de Box e Jenkins para a modelagem de séries temporais é a identificação. De acordo com Morettin e Toloi (2006), para a identificação de potenciais modelos ARIMA a serem ajustados aos dados, as principais fontes de informação são as funções de autocorrelações e autocorrelações parciais amostrais.

Os mesmos autores definem três principais passos para o procedimento de identificação dos modelos, que têm como objetivo a determinação dos valores dos parâmetros p , d e q (além de P , D e Q no caso de séries com componente sazonal).:

- i. Verificar se há a necessidade de uma transformação na série original, para estabilização da variância. Essa verificação é realizada visualmente, através da plotagem dos valores.
- ii. Diferenciar a série, conforme explicado nas Seções 2.3.4 e 2.3.5, até que seja obtido um processo estacionário, para que seja possível a aplicação de um modelo $ARMA(p, q)$ à série.
- iii. Identificar qual é o modelo $ARMA(p, q)$ com comportamento semelhante ao da série, a partir da comparação das funções de autocorrelações e autocorrelações parciais dos dados com as autocorrelações e autocorrelações parciais teóricas de modelos AR, MA e ARMA, com ou sem componente sazonal.

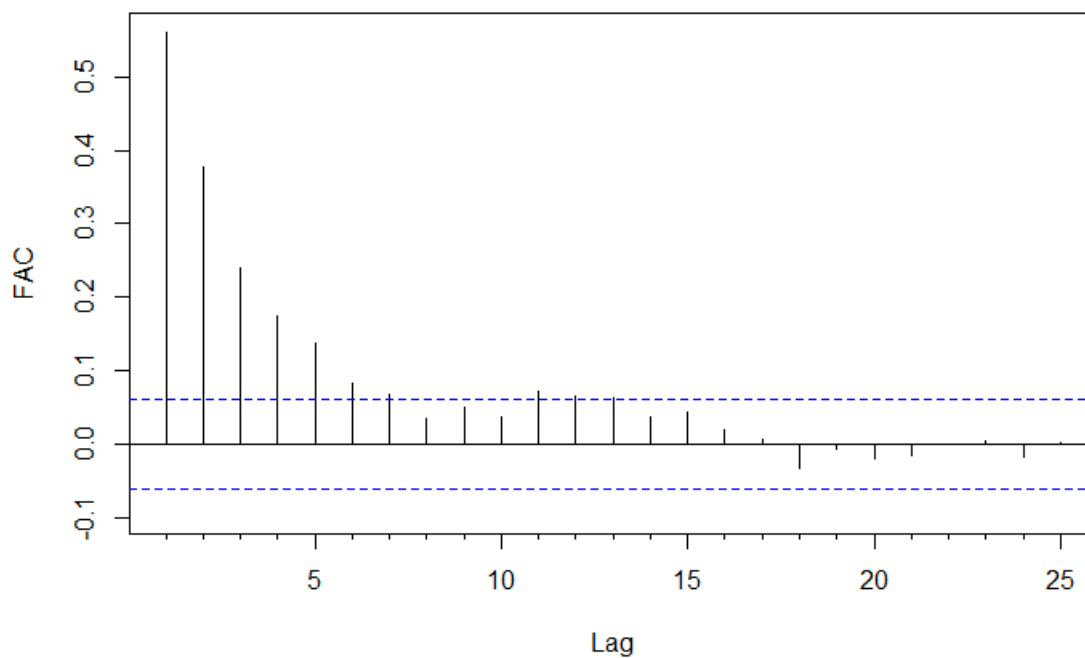
Para o cálculo da fac e $facp$, é proposto que seja utilizado algum software estatístico, que calcule as funções para todos os *lags* desejados. Para séries sazonais mensais, como a que será utilizada no caso real, sugere-se que as funções sejam analisadas ao menos até o *lag* 24, de modo a conseguir verificar dois períodos sazonais inteiros.

Em seguida, é proposto que a fac e a $facp$ da série sejam analisadas graficamente, juntamente com os valores críticos para a significância dos valores, dados por $\pm 2/\sqrt{n}$. A partir da identificação das fac e $facp$ significativas, será possível determinar alguns modelos candidatos, de acordo com o comportamento teórico dessas funções para cada tipo de modelo. Os principais comportamentos de séries temporais autorregressivas e de médias móveis foram apresentados ao longo da Seção 2.6 do presente trabalho.

Tomando como exemplo um processo $AR(p)$ teórico, relembremos que a fac decai de acordo com exponenciais ou senoides amortecidas e a $facp$ é significativa apenas para *lags* $k \leq p$. Para visualizar este comportamento na prática, utilizou-se a função “*arima.sim*” da biblioteca “*stats*” do RStudio, foram simuladas 1000 observações de um processo $AR(2)$, com coeficientes $\phi_1 = 0.5$ e $\phi_2 = 0.1$, estacionário. No mesmo programa, através das funções “*Acf()*” e “*Pacf()*” da biblioteca “*forecast*”, foram plotados os gráficos da fac e da $facp$, respectivamente, para este processo. Os

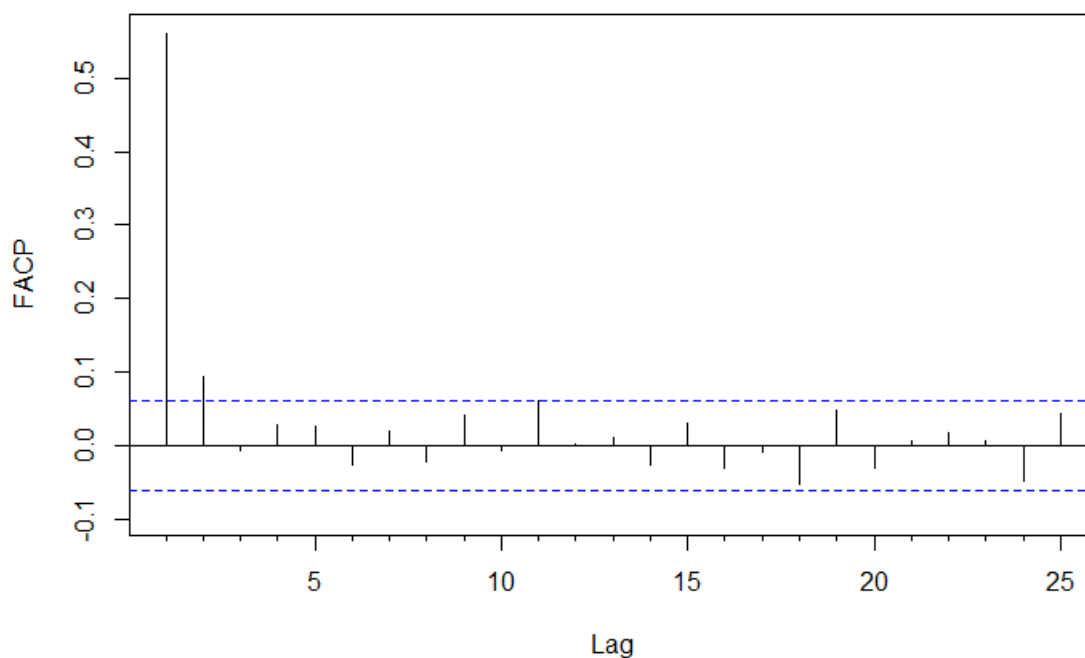
resultados obtidos estão expostos nas Figuras 9 e 10, juntamente com os valores críticos das funções, sinalizados pela linha azul pontilhada.

Figura 9 - FAC de um processo AR(2) puro



Fonte: Elaborado pelo autor.

Figura 10 - FACP de um processo AR(2) puro



Fonte: Elaborado pelo autor.

Como podemos notar, o comportamento deste processo segue de maneira muito próxima o comportamento de um processo AR(2) teórico, como era de se esperar. A fac decai exponencialmente, sendo significativa até *lags* mais altos. Enquanto isso, a

facp tem um rápido decaimento para 0, com apenas os primeiros coeficientes mostrando-se significativos.

Na prática, é muito difícil que um modelo puramente autorregressivo ou de médias móveis seja o mais eficaz para a descrição de uma série temporal, a não ser com a utilização de dezenas de parâmetros. Dessa maneira, de modo a priorizar modelos mais parcimoniosos, é mais comum que seja utilizada a abordagem ARMA em detrimento de uma abordagem AR ou MA pura.

4.3. Estimação e seleção do modelo

Após a etapa de identificação, ocorre a etapa de estimação, durante a qual os parâmetros do modelo identificado são estimados. Ao longo dela, é utilizado um procedimento iterativo de estimação não-linear de mínimos quadrados.

Consideremos um modelo ARIMA(p, d, q) cujos $p + q + 1$ parâmetros podem ser colocados no vetor $\xi = (\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$, onde $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)$ e $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)$. No caso em que $d > 0$, é suposto que $\mu_W = 0$. Caso contrário, μ é considerado como mais um parâmetro a ser estimado, totalizando $p + q + 2$ parâmetros.

Um dos métodos mais utilizados para a estimação dos parâmetros de modelos ARIMA é o da máxima verossimilhança. Assim, para estimar ξ , dadas as N observações $Z_1, \dots, Z_N$, iremos considerar a função de verossimilhança $L(\xi|Z_1, \dots, Z_N)$ como função de ξ . Os estimadores de máxima verossimilhança (EMV) serão os valores do vetor ξ que maximizarem a função L , ou $l = \log L$.

Uma das suposições do método é de que o processo a_t é gaussiano, ou seja, $a_t \sim N(0, \sigma_a^2)$ para cada t . Nesse caso, os EMV assumirão valores semelhantes aos dos estimadores de mínimos quadrados (EMQ).

Quando há a necessidade de se tomar d diferenças para atingir a estacionariedade do processo, restam $n = N - d$ observações $W_1, \dots, W_N$, onde $W_t = \Delta^d Z_t$. Dado que o modelo ARMA(p, q) resultante é estacionário e invertível, temos que:

$$a_t = \tilde{W}_t - \phi_1 \tilde{W}_{t-1} - \dots - \phi_p \tilde{W}_{t-p} + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q}, \quad (4.3)$$

onde $\tilde{W}_t = W_t - \mu_W$.

Para que possamos calcular os a_t , é necessário que obtenhamos os valores iniciais para os $\tilde{W}_t$ e para os a_t . Esses valores iniciais podem ser obtidos através de um procedimento condicional, onde são utilizadas estimativas consideradas razoáveis para o cálculo deles.

Lembremos que os a_t possuem distribuição normal. Assim, a função de densidade conjunta de $a_1, \dots, a_n$ é dada por:

$$f(a_1, \dots, a_n) = (2\pi)^{-n/2} (\sigma_a)^{-n} \exp \left\{ - \sum_{t=1}^n \frac{a_t^2}{2\sigma_a^2} \right\}. \quad (4.4)$$

Com isso, a função de verossimilhança é obtida da seguinte maneira:

$$\begin{aligned} L(\xi|W_t, a_t) \\ = (2\pi)^{-n/2} (\sigma_a)^{-n} \exp \left\{ - \frac{1}{2\sigma_a^2} \sum_{t=1}^n (\tilde{W}_t - \phi_1 \tilde{W}_{t-1} - \dots - \phi_p \tilde{W}_{t-p} + \theta_1 a_{t-1} + \dots \right. \\ \left. + \theta_q a_{t-q})^2 \right\}. \end{aligned} \quad (4.5)$$

Tomando-se o logaritmo da função, temos:

$$l(\xi|W_t, a_t) \propto -n \log \sigma_a - \frac{S(\eta|W_t, a_t)}{2\sigma_a^2}, \quad (4.6)$$

onde

$$S(\eta|W_t, a_t) = \sum_{t=1}^n a_t^2 (\eta|W_t, a_t), \quad (4.7)$$

é chamada de soma de quadrados condicional. Os estimadores iniciais dos parâmetros ξ serão aqueles que maximizarem a função $l(\xi|W_t, a_t)$, o que é equivalente a minimizar a função $S(\eta|W_t, a_t)$.

É importante ressaltar que dificilmente a fac e a facp de séries reais terão valores claramente aderentes a um dos modelos teóricos estudados, assumindo um comportamento mais complexo. Dessa maneira, o presente trabalho propõe que uma variedade de modelos candidatos seja testada, e que todos eles sejam comparados a partir de um mesmo critério quantitativo, logo após a fase de estimação.

Há diversos métodos quantitativos que podem ser utilizados para a seleção do modelo, como aqueles baseados em funções penalizadoras, ou seja, funções que buscam penalizar modelos proporcionalmente ao número de parâmetros que o compõem. De acordo com Morettin e Toloí (2006), estes métodos têm como objetivo a minimização de uma função penalizadora, que pode assumir diversas formas de acordo com o método utilizado.

Um dos métodos mais difundidos atualmente para tal, que será utilizado no presente trabalho, é o Critério de Informação de Akaike (AIC), proposto por Akaike (1974). Neste caso, a função que deve ser minimizada é a seguinte:

$$AIC(k, d, l) = N \ln \hat{\sigma}_a^2 + \frac{N}{N-d} 2(k + l + 1 + \delta_{d0}) + N \ln 2\pi + N, \quad (4.8)$$

em que $\delta_{d0} = 1$ se $d = 0$ e $\delta_{d0} = 0$ caso contrário; $\hat{\sigma}_a^2$ é o EMV de σ_a^2 , que é a variância residual obtida ao ajustar um modelo ARIMA(k, d, l) às N observações de uma série temporal.

Dessa maneira, quanto maior o número de parâmetros, maior é o termo penalizador, o que resulta na priorização de modelos mais parcimoniosos. A seleção de modelos através deste critério é baseada no cálculo do AIC de diversos modelos candidatos e na seleção daquele que obtiver o menor resultado para esta estatística.

Com a definição do melhor modelo para a descrição da série, resta testar se todos os coeficientes que o compõem são significativos. Caso contrário, os coeficientes que não o são devem ser retirados do modelo, e ele deve ser reestimado apenas com os parâmetros restantes. Assim, é necessário avaliá-lo novamente a partir do critério AIC, que será comparado com a avaliação dos demais modelos candidatos. O modelo escolhido, que será levado para a fase seguinte, de diagnóstico, é aquele com menor AIC cujos coeficientes são todos significativos ao nível de 5%.

4.4. Diagnóstico do modelo

Como apresentado no Capítulo 2, a última etapa da abordagem Box e Jenkins para a modelagem de séries temporais é o diagnóstico. A partir do modelo obtido na fase de identificação e posteriormente estimado, iremos à fase final, para verificar se o modelo é adequado para descrever a série temporal em questão. Para tal, serão analisados os resíduos a_t do modelo, que são dados pela diferença entre os valores dos dados e os valores estimados:

$$a_t = Z_t - \hat{Z}_t,$$

onde $\hat{Z}_t$ são os valores estimados pelo modelo para as observações Z_t . Para que o modelo seja classificado como satisfatório, é necessário que os resíduos a_t sejam ruído branco.

Dessa maneira, os testes de diagnóstico serão realizados com o intuito de testar se as propriedades de a_t são aderentes àquelas do ruído branco. Três propriedades essenciais que os resíduos devem ter para que o modelo em questão seja considerado aderente aos dados são:

- i. Os resíduos não são autocorrelacionados;

- ii. A média dos resíduos é igual a zero;
- iii. Os resíduos seguem uma distribuição normal com variância constante.

Para testar a primeira propriedade, dois testes serão realizados. O primeiro deles é o gráfico, a partir da plotagem da $\hat{a}_t$ e $\hat{a}_{t+h}$ amostral dos resíduos. A análise gráfica é muito semelhante à realizada na etapa de identificação do modelo, com a análise de ambos os gráficos e verificação das autocorrelações e autocorrelações parciais significativas do modelo. Nesta etapa, entretanto, o ideal é que nenhuma delas seja significativa, o que evidencia que os resíduos não são autocorrelacionados.

Outro teste muito utilizado nesta etapa e que será aplicado no presente trabalho é o de Ljung-Box, proposto por Ljung e Box (1978) a partir do teste de Box e Pierce (1970). De acordo com Morettin e Toloí (2006), este teste, apesar de não detectar quebras específicas no comportamento do ruído branco, pode demonstrar se os valores das autocorrelações dos resíduos são muito elevados.

A hipótese nula deste teste é de que os resíduos do modelo são independentes. A estatística utilizada para tal é a seguinte:

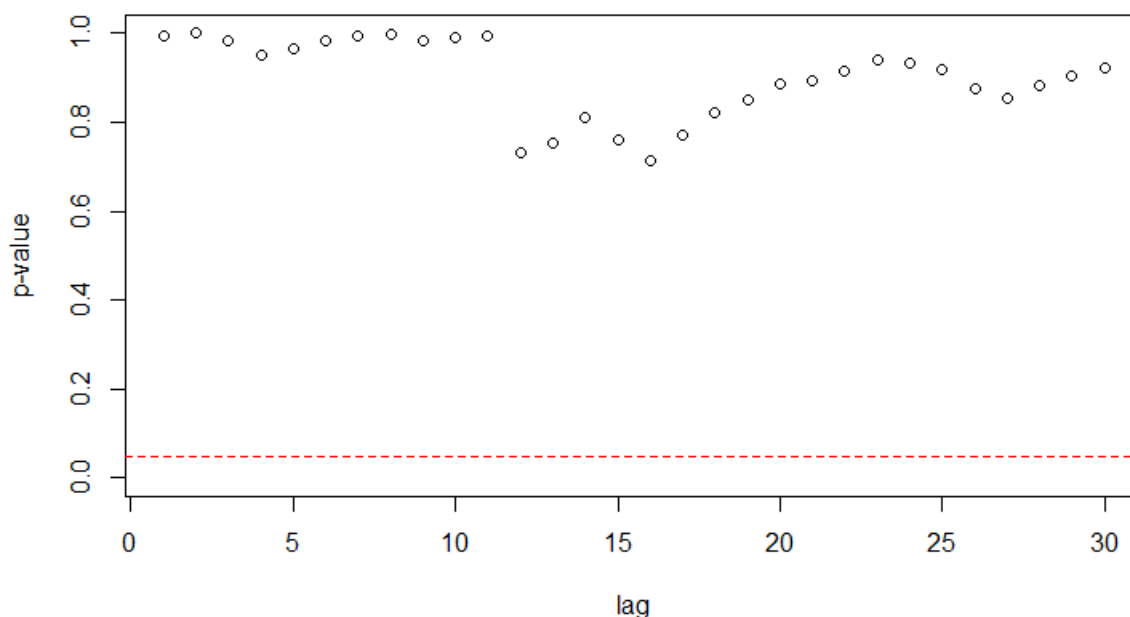
$$Q(K) = n(n+2) \sum_{j=1}^K \frac{\hat{r}_j^2}{N-j}, \quad (4.9)$$

onde $\hat{r}_j$ são as estimativas das autocorrelações dos resíduos. A estatística terá aproximadamente uma distribuição χ^2 com $K - p - q$ graus de liberdade, onde $p + q$ é o número de parâmetros do modelo ARMA estimado. A hipótese de independência, ou de ausência de autocorrelação, é rejeitada para valores grandes de Q .

O nível de significância α a ser utilizado no presente trabalho será de 5%. Dessa maneira, para p-valores acima de 5%, não será possível rejeitar a hipótese nula, assumindo os resíduos a_t como independentes.

Além desta estatística, o teste também será analisado graficamente, a partir da análise dos p-valores para cada um dos *lags* entre as observações. Assim, para cada *lag* τ , um p-valor menor do que o nível de significância α indica que a hipótese nula de que os resíduos que distam τ observações entre si são independentes deve ser rejeitada. Para que seja assumida independência dos resíduos, é necessário que todos os pontos do gráfico estejam acima da linha de significância, como o caso mostrado na Figura 11, de um modelo ARIMA ajustado para uma série AR(2) gerada pela função “arima.sim”. O gráfico em questão foi gerado pela função “LBQPlot” da biblioteca “FitAR” do RStudio.

Figura 11 - Gráfico com os p-valores de cada *lag* do teste Ljung-Box para o modelo ajustado ao processo AR(2)



Fonte: Elaborado pelo autor.

Já para o teste de normalidade, serão aplicadas aos resíduos a_t as duas ferramentas apresentadas na Seção 2.4: o gráfico QQ e o teste Shapiro-Wilk. A partir deles, é possível concluir se $a_t \sim N(0, \sigma^2)$, onde σ^2 é a variância dos resíduos do modelo. Novamente, uma das ferramentas utilizadas é gráfica e outra é estatística, o que confere uma confiança adicional para a aceitação ou não da normalidade dos resíduos. Para o teste de Shapiro e Wilk (1965), também será utilizado um nível de significância de 5%. Assim, se o teste retornar um p-valor superior a 0,05, a hipótese nula de normalidade dos resíduos não poderá ser rejeitada, e poderemos considerá-los como aderentes à distribuição Gaussiana.

Com o modelo em questão tendo resultados favoráveis nos testes de independência (análise gráfica da fac e da facp e teste Ljung-Box) e de normalidade (gráfico QQ e teste Shapiro-Wilk) dos resíduos, ele pode ser validado e considerado adequado para seguir às próximas etapas, de previsão e controle estatístico. Caso contrário, os resíduos não podem ser considerados ruído branco e conclui-se que o modelo não é satisfatório para descrever a série temporal. Nesse caso, é necessário fazer a seleção de outro modelo na fase da identificação, estimá-lo e realizar novamente o diagnóstico, até que seja encontrado um cujos resíduos satisfaçam às condições necessárias.

4.5. Determinação dos limites de controle

O primeiro passo para a determinação dos limites de controle é a seleção do gráfico que será utilizado, conforme o fluxograma da Figura 6. Dado que o processo a ser modelado é uma série temporal, será utilizada uma das abordagens encontradas no ramo da direita, já que processos modelados por séries temporais possuem autocorrelações significativas.

Dentre as opções sugeridas por Montgomery (2012) para esta situação, a primeira é a mais propícia à situação do presente trabalho, que se propõe a ajustar um modelo SARIMA aos dados do processo. Dessa maneira, resta decidir quais dos gráficos de controle tradicionais serão aplicados ao processo em si para estabelecer os limites de controle da série.

De acordo com Knoth e Schmid (2004), para séries autocorrelacionadas são recomendados os gráficos de controle CUSUM e EWMA, ambos com memória. Como os dados de processos dessa natureza são dependentes de observações passadas, é esperado que gráficos de controle com memória desempenhem melhor nessa situação. Outro argumento a favor dos gráficos CUSUM e EWMA, neste caso, é de que eles são mais eficientes em detectar pequenas variações no processo.

Levando em consideração a facilidade de uso e de interpretação, optou-se por seguir com o gráfico EWMA para realizar o controle estatístico da série econômica no presente trabalho.

Para definir os limites de controle, é necessário selecionar um valor alvo para o ARL, ou o comprimento médio da sequência antes que uma observação seja sinalizada como fora de controle. Conforme Montgomery (2012), em gráficos de controle de Shewhart para $\bar{x}$, com limites de três desvios padrões, a probabilidade p de que um ponto caia fora dos limites com o processo em controle é de 0,0027, com um ARL de 370.

Entretanto, no presente trabalho, a série a ser utilizada possui periodicidade mensal, tendo uma frequência muito menor em comparação com as aplicações tradicionais de gráficos Shewhart para $\bar{x}$. No caso de uma série mensal, um ARL de 370 indicaria uma observação fora de controle a cada 31 anos, prejudicando a capacidade do gráfico de detectar observações que realmente estão fora de controle. Assim, foi escolhido um ARL_{alvo} menor, de 36, o que indicaria uma observação fora de controle a cada três anos.

Os limites de controle do gráfico EWMA serão definidos a partir de simulações, com base no valor alvo para o ARL. Para tal, será selecionado um valor inicial para os limites de controle (a partir do desvio padrão σ dos resíduos do modelo, multiplicado por um fator L), e a série em questão será simulada N vezes. A série simulada será gerada a partir dos coeficientes do modelo SARIMA e da simulação de seus resíduos $a_t \sim N(0, \sigma^2)$. Em seguida, os valores de Z_t serão calculados, a partir da equação (3.22), e comparados aos limites de controle para verificar se estão dentro ou fora do intervalo definido por eles, até que seja encontrado o primeiro ponto fora de controle – contabilizando, assim o tamanho da sequência.

Após a realização de todas as N simulações, o tamanho médio das sequências será calculado. Caso este valor esteja próximo ao alvo, os limites de controle são fixados. Caso contrário, aumenta-se o valor de L e realizam-se novamente as simulações, finalizando o processo quando o valor do ARL estiver próximo de ARL_{alvo} .

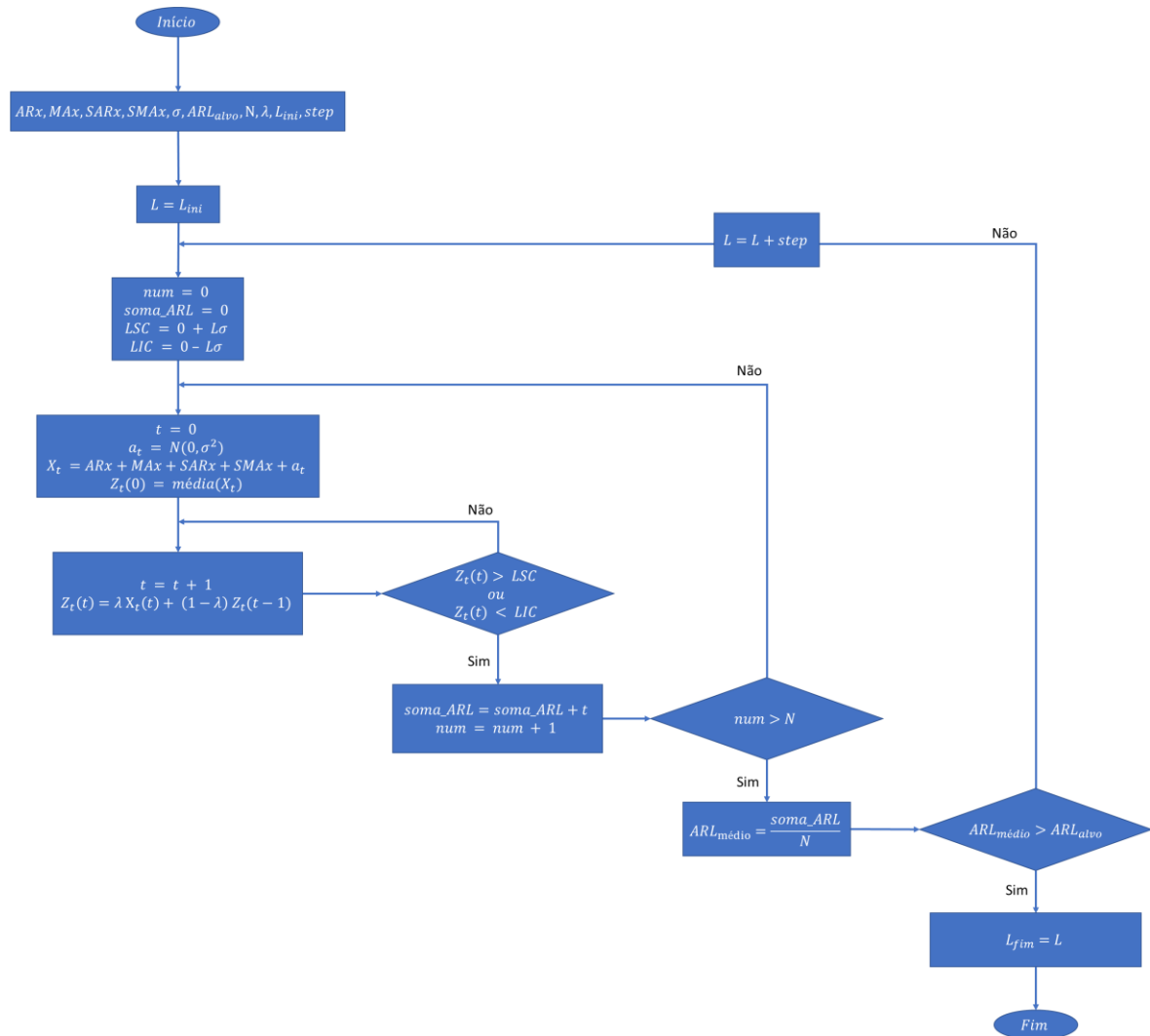
Antes da realizar as simulações, porém, ainda é necessário definir o parâmetro λ , utilizado para ponderar os pesos da observação atual e das anteriores. No presente trabalho, será utilizado um valor intermediário de 0,6 para λ , de maneira a manter um meio termo entre a detecção de pequenas variações na média e a ocorrência de alarmes falsos.

Para apresentar o algoritmo utilizado para definir os limites de controle de forma mais ilustrativa, foi elaborado o fluxograma apresentado na Figura 12. A lista a seguir contém uma breve explicação acerca das variáveis utilizadas no algoritmo que ainda não foram especificadas:

- ARx : componente autorregressivo do modelo estimado;
- $SARx$: componente autorregressivo sazonal do modelo estimado;
- MAx : componente de médias móveis do modelo estimado;
- $SMAx$: componente de médias móveis sazonal do modelo estimado;
- L_{ini} : valor inicial do fator que determina os limites de controle;
- $step$: valor de incremento do fator multiplicador do limite de controle;
- L_{fim} : valor final do fator multiplicador do limite de controle;
- a_t : série de valores simulados para os ruídos brancos da série temporal;
- X_t : série de valores estimados da série temporal original;
- Z_t : série de valores da estatística utilizada para o monitoramento através do gráfico EWMA;

- num = contador do número de iterações que foram realizadas;
- $soma_ARL$ = somatória dos tamanhos de sequências (*run lengths*) das iterações para um mesmo valor de L ;
- $ARL_{médio}$ = média dos *run lengths* de cada iteração para um mesmo valor de L .

Figura 12 - Fluxograma do algoritmo para cálculo dos limites de controle



Fonte: Elaborado pelo autor.

4.6. Obtenção dos ARL fora de controle

Após a definição dos limites de controle, é necessário verificar a capacidade do gráfico em identificar processos fora de controle. Assim, iremos avaliar o ARL_1 do gráfico EWMA, ou seja, quantas observações o gráfico demora para identificar que o processo controlado por ele está fora de controle, quando ele realmente está fora de controle.

Portanto, é necessário simular um processo sob a hipótese H_1 similar ao realizado na Seção 4.5. Para simular este processo fora de controle, serão utilizados os coeficientes do modelo SARIMA ajustado à série temporal com resíduos a_t aderentes à distribuição $N(0, k\sigma^2)$, onde k é um fator que será variado de 1 a 5 para alterar a variância dos resíduos gerados. A expectativa é que, à medida que o valor de k aumente, o ARL do gráfico de controle diminua, indicando que ele é eficiente em detectar processos fora de controle.

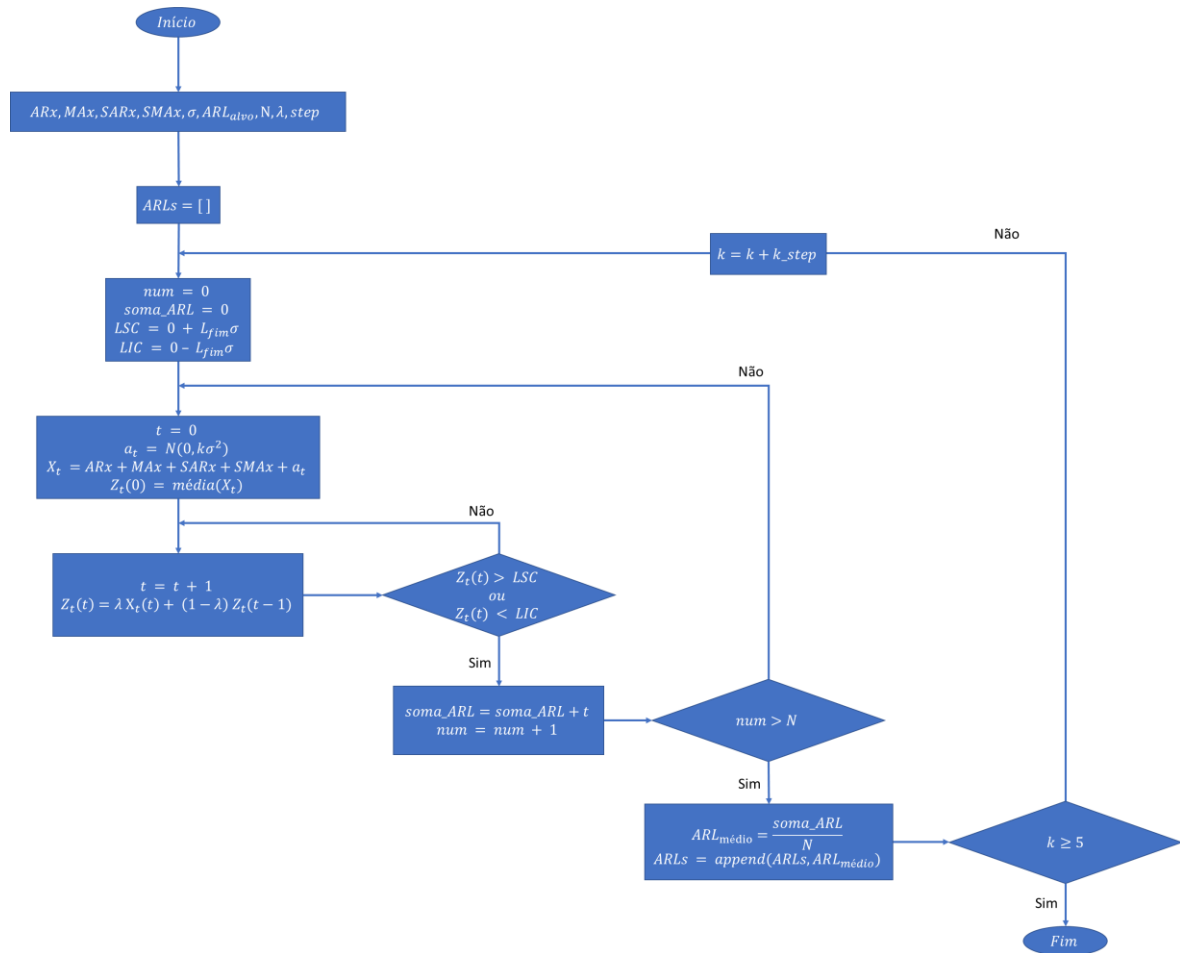
Assim, os valores dos limites de controle serão fixados em $0 \pm L_{fim}\sigma$, de acordo com o valor de L_{fim} encontrado na etapa anterior. Da mesma forma, o parâmetro λ também assumirá o mesmo valor utilizado para o cálculo dos limites de controle, de 0,2.

O algoritmo utilizado para a obtenção dos ARL_1 é muito semelhante ao utilizado para o cálculo dos limites de controle. A principal diferença entre eles é que, para o cálculo dos limites de controle, variava-se o valor de L a cada iteração, até que o ARL médio das simulações assumisse um valor maior do que ARL_{alvo} . Agora, a cada iteração, será variado o valor do coeficiente k , de maneira a calcular o ARL de processos com variâncias distintas. Este algoritmo está exibido no fluxograma da Figura 13.

Fora as variáveis já apresentadas para o algoritmo de cálculo dos limites de controle, serão utilizadas as seguintes variáveis adicionais:

- $ARLs$: lista com os valores dos ARLs para cada um dos k testados;
- k_step : valor de incremento para o fator k .

Figura 13 - Fluxograma do algoritmo para cálculo dos ARLs fora de controle



Fonte: Elaborado pelo autor.

4.7. Avaliação dos resultados

A partir da modelagem da série temporal, será possível construir previsões para as observações futuras da série, buscando identificar como ela deve evoluir alguns períodos para frente, o que pode ser extremamente útil para a tomada de decisão no presente.

No presente trabalho, propõe-se que sejam utilizados de 10% a 20% dos dados disponíveis para a testagem da eficácia dos modelos como método de previsão, com as demais observações sendo efetivamente utilizadas para a modelagem da série, sendo submetidas aos passos da abordagem Box e Jenkins. O conjunto de dados utilizado para a previsão será chamado de conjunto de teste, ao passo que os demais serão denominados conjunto de treinamento.

A vantagem de separar os dados nestes dois conjuntos é ter a possibilidade de aplicar o modelo em dados que não foram utilizados para a sua criação, testando na prática se ele é realmente eficaz para modelar dados novos, como seria em um caso

real para a previsão de valores futuros. Utilizando os mesmos dados para o treinamento e o teste, é esperado que o modelo tenha um bom desempenho. Assim, não é possível ter uma estimativa fidedigna de qual será o desempenho do modelo na prática.

Para dados sazonais com periodicidade mensal, como é o caso do presente trabalho, é recomendado que o conjunto de teste tenha ao menos 12 observações, de modo que seja possível realizar as previsões ao longo de um período inteiro.

Neste trabalho, é proposto que, após as etapas de identificação, estimação e diagnóstico (realizadas apenas nos dados de treinamento), o modelo encontrado seja utilizado para realizar a previsão de alguns períodos a frente, e que essas previsões sejam comparadas aos valores reais do conjunto de teste para que seja avaliada a performance do modelo. O critério a ser utilizado nesta avaliação será a raiz do erro quadrático médio (RMSE, da sigla em inglês para *root mean square error*). Dados um processo Z_t , $t \in T$, e suas previsões $\hat{Z}_t$, realizadas para $t = h, h + 1, \dots, h + n$, o RMSE é dado por:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=h}^{h+n} (\hat{Z}_t - Z_t)^2}{(h+n) - h}}. \quad (4.10)$$

Assim, podemos notar que quanto menor o RMSE, melhor é o modelo para realizar previsões da série em questão. Este critério pode ser utilizado em conjunto com o AIC para a seleção entre dois ou mais modelos, dado que ambos são parâmetros para a avaliação da performance deles.

Por fim, para avaliar o desempenho do modelo de previsão, serão calculados os intervalos de confiança de 90% das previsões, de modo a verificar se as observações de teste se encontram dentro deles. Espera-se, portanto, que aproximadamente 90% das observações fiquem dentro deste intervalo, dado o nível de confiança selecionado.

Os gráficos de controle serão utilizados para monitorar a série temporal modelada, de maneira a indicar se o processo em questão está dentro ou fora de controle. Em um processo industrial, o indicativo de que o processo está fora de controle é muito importante, por apontar a necessidade de intervenção. No caso do índice sobre o qual será aplicada esta metodologia no Capítulo 5, o IBC-Br, este indicativo servirá para apontar os períodos em que a atividade econômica brasileira está acima ou abaixo do nível esperado – dadas as informações das observações passadas –, o que pode

servir como um insumo importante no processo de tomada de decisão para a alocação de ativos relacionados à economia do Brasil, como títulos do Tesouro Direto e índices de renda variável, como o próprio Ibovespa.

5. APLICAÇÃO EM CASO REAL

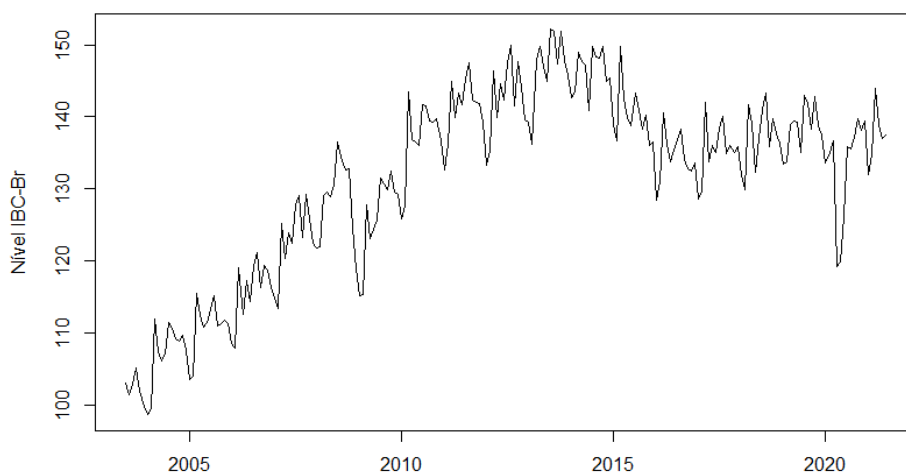
Neste capítulo, a metodologia apresentada no Capítulo 4, desenvolvida a partir da revisão bibliográfica de séries temporais e CEP, será aplicada a um caso real. A série selecionada para a aplicação desta metodologia foi a do IBC-Br, conforme explicitado no capítulo introdutório. Assim, ao longo deste capítulo, será demonstrado como foi realizada a modelagem desta série e quais foram os resultados obtidos com a previsão e com o controle estatístico dela através de gráficos de controle. Por fim, será feita uma análise destes resultados, de modo a determinar a eficácia da metodologia em atingir os objetivos propostos pelo trabalho.

5.1. Coleta e ajuste dos dados

Os dados do IBC-Br foram extraídos de fontes do próprio BC, a instituição responsável pelo cálculo e divulgação desta estatística. A série temporal foi obtida através do site do Sistema Gerenciador de Séries Temporais do banco, tendo sido acessada pelo seu API (Interface de Programação de Aplicações, da sigla em inglês) através do RStudio.

À época em que o sistema foi acessado, o último dado divulgado havia sido o referente ao mês de junho de 2021. A amostra utilizada para o estudo é composta pelos dados a partir de julho de 2003, totalizando 18 períodos completos, o equivalente a 216 observações. A série original, sem transformações, é expressa em nível, e a partir de agora ela será referida no presente trabalho como X_t . Na Figura 14, podemos ver seu gráfico.

Figura 14 - Série temporal do IBC-Br

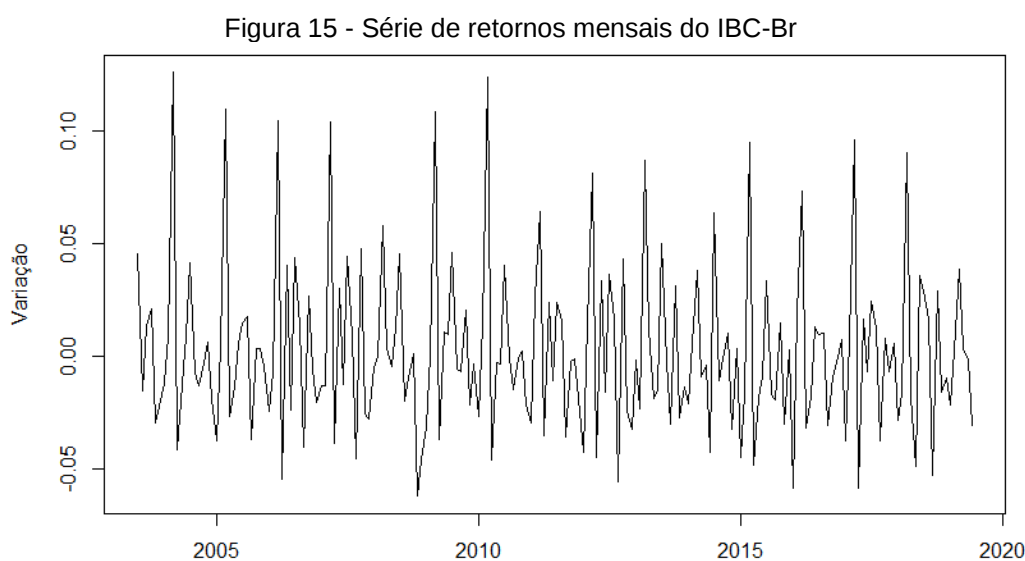


Fonte: Elaborado pelo autor.

Para que seja possível avaliar a precisão do modelo de previsão e do gráfico de controle, os últimos dois anos de dados foram retirados da série e serão utilizados para teste, de modo que apenas os dados de julho de 2003 a junho de 2019 serão utilizados para a construção do modelo. De qualquer maneira, a quantidade de dados é suficiente para uma modelagem ARIMA eficiente, conforme o critério levantado na Seção 4.1. Assim, é necessário analisar as propriedades da série ao longo deste período, de modo a verificar se ela é adequada para os modelos propostos.

Visualmente, como mostra a Figura 14, fica evidente que a série X_t não é estacionária. Sua média varia bastante ao longo dos períodos, e há tendências locais. De 2003 a 2014, é possível notar uma tendência quase linear de crescimento, que se reverte para uma tendência de queda entre 2014 e 2016. Desde 2016, a série mostra uma leve tendência de alta, com uma queda fora do padrão no primeiro semestre de 2020, devido ao início da pandemia de COVID-19. O teste ADF, realizado através da função “adf.test” da biblioteca “tseries” do RStudio, apenas confirma que a série em questão não pode ser considerada estacionária (dada a hipótese nula de que a série não é estacionária e considerando um α de 5%), retornando um p-valor de 0,46.

Dessa maneira, torna-se necessário empregar alternativas para buscar a estacionariedade de X_t . Para tal, foi calculada a série de variações mensais dela, R_t , exibida na Figura 15.

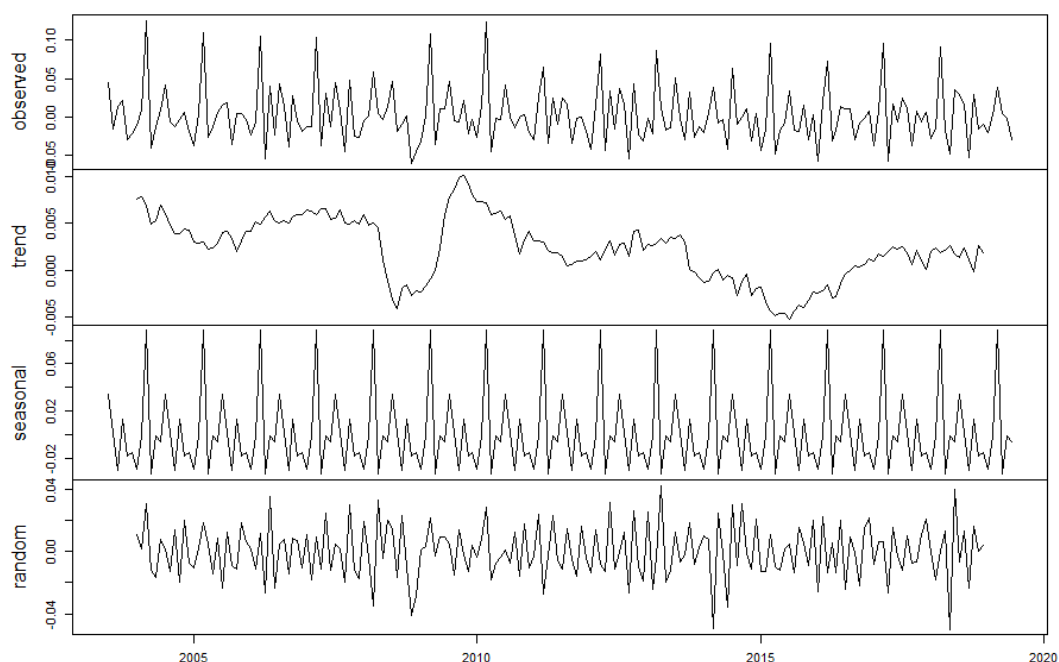


Fonte: Elaborado pelo autor.

Conforme a Figura 9, que mostra a análise gráfica de R_t , a série parece ser estacionária, com uma média constante, próxima de 0, ao longo de toda a sua extensão, e sem tendências claras. Entretanto, é possível notar picos e vales

equidistantes ao longo da série, separados por 12 períodos. Esse comportamento é um forte indício de sazonalidade, indicando que alguns meses tendem a ter valores mais altos ou mais baixos que a média da série. Para auxiliar na análise da sazonalidade, foi aplicada a função “decompose”, do pacote “stats”, sobre a série R_t , que decompõe a série temporal em três componentes: tendência, sazonalidade e aleatório. Para que uma série possa ser modelada de maneira mais adequada, é esperado que o componente aleatório seja predominante. Na Figura 16, está ilustrada a decomposição da série R_t .

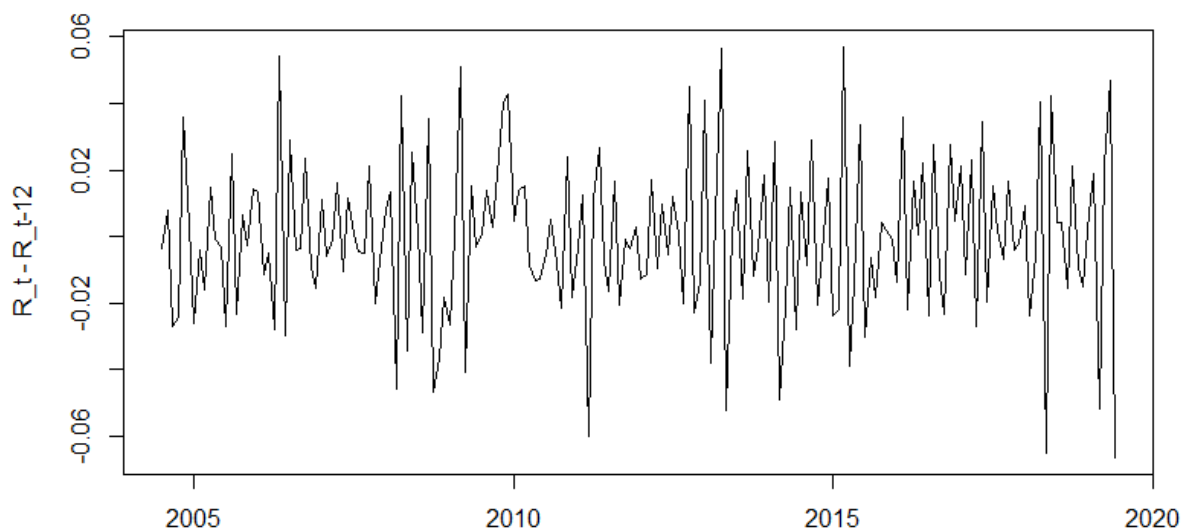
Figura 16 - Decomposição de R_t



Fonte: Elaborado pelo autor.

A plotagem da série decomposta evidencia que realmente a sazonalidade tem um papel relevante na série, com a maior parte dos componentes sazonais variando entre -0,02 e 0,08, tendo valores absolutos maiores, em média, do que o componente aleatório, que varia principalmente entre -0,04 e 0,04. Assim, é necessário aplicar uma transformação sobre a série de forma a buscar remover este componente sazonal, para que seja possível modelar sua parte aleatória. Para tal, foi calculada a primeira diferença sazonal de R_t , obtendo-se a série D_t , mostrada na Figura 17.

Figura 17 - Série de retornos do IBC-Br diferenciada sazonalmente

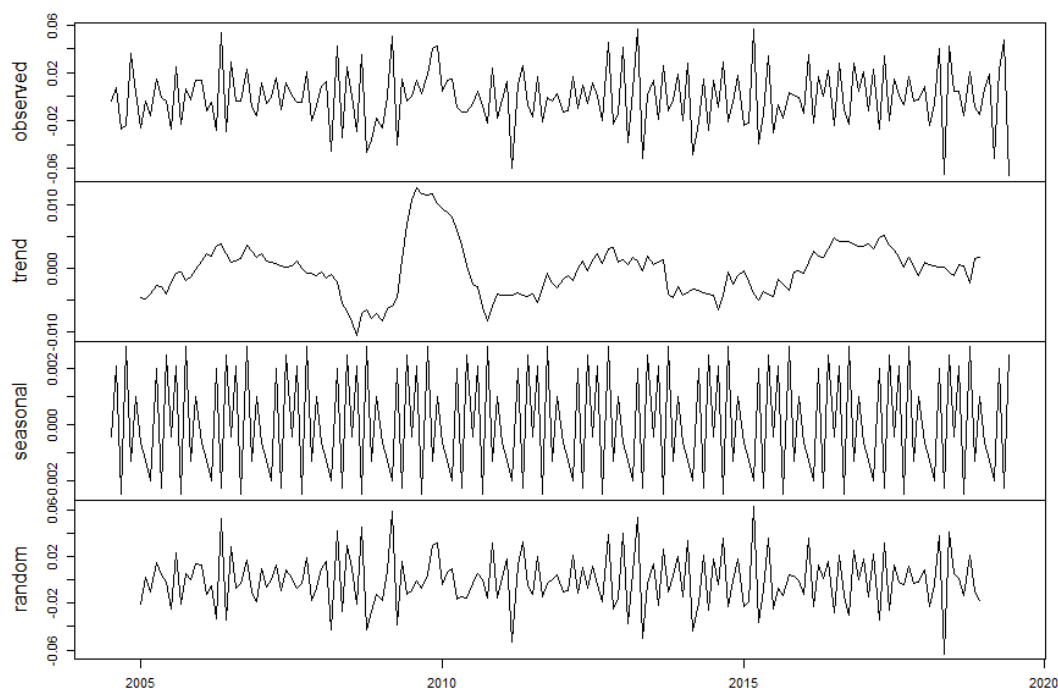


Fonte: Elaborado pelo autor.

À primeira vista, a série D_t não aparenta ter nenhum problema, com uma média constante ao redor de 0 e sem demonstrar tendências e sazonalidade. Aplicando o teste ADF sobre ela, o valor retornado para o p-valor é menor do que 0,01, o que nos permite rejeitar a hipótese nula de não estacionariedade.

Realizando a decomposição da série, podemos notar que o componente aleatório de D_t é predominante, sendo adequada para a modelagem. A Figura 18 ilustra a decomposição da série.

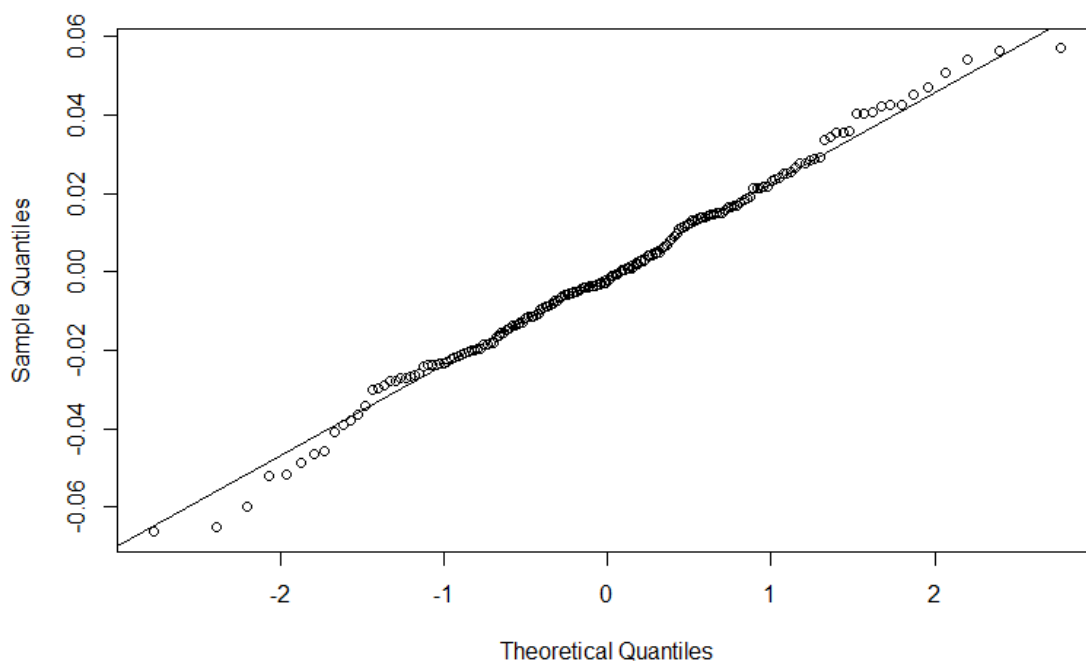
Figura 18 - Decomposição de D_t



Fonte: Elaborado pelo autor.

Por fim, foi realizada uma análise da distribuição de D_t , de modo a verificar se a série é aderente à distribuição normal. Assim, foi elaborado o gráfico QQ, que pode ser visto na Figura 19, para a análise visual da normalidade.

Figura 19 - Gráfico QQ da série D_t

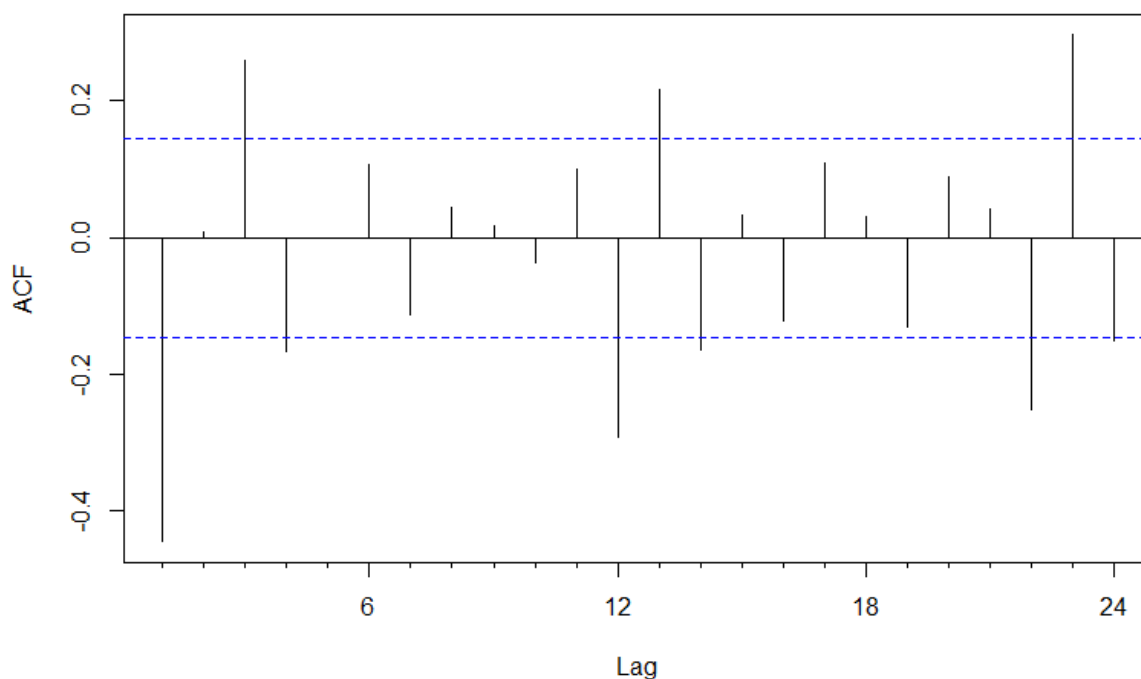


Fonte: Elaborado pelo autor.

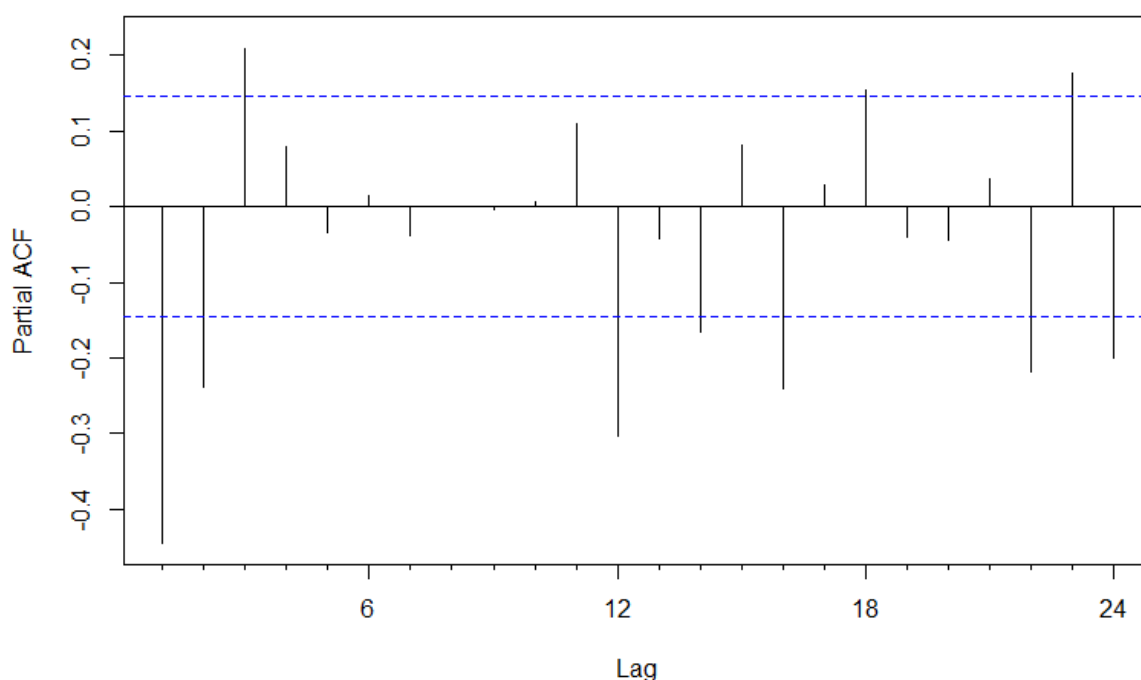
A maior parte dos pontos encontra-se próxima da reta normal, o que indica que a série deve ser aderente a esta distribuição. No entanto, nas extremidades há alguns pontos mais afastados da reta. Dessa maneira, para verificar estatisticamente se a distribuição de D_t é gaussiana, foi aplicado o teste de Shapiro-Wilk sobre ela, que retornou um p-valor de 0,54. Assim, é possível assumir a hipótese nula de normalidade como verdadeira, dado o nível de significância de 5%.

5.2. Identificação do modelo

O primeiro passo para a identificação do modelo é a análise das autocorrelações e autocorrelações parciais da amostra. Assim, através do RStudio, foram plotados os gráficos das duas funções, que foram analisados visualmente. Estes gráficos estão apresentados nas Figuras 20 e 21, respectivamente. As linhas azuis pontilhadas, presentes em ambos os gráficos, representam os limites para as fac e $facp$ positivas, iguais a aproximadamente $\pm 0,15$ neste caso.

Figura 20 - Gráfico de autocorrelação da série D_t 

Fonte: Elaborado pelo autor.

Figura 21 - Gráfico de autocorrelação parcial da série D_t 

Fonte: Elaborado pelo autor.

Podemos notar que os valores significativos da fac e da $facp$ estão concentrados principalmente nos $lags$ mais baixos e nos $lags$ próximos de múltiplos de 12, o que indica que ainda deve existir um componente sazonal presente nos dados. Além disso, percebe-se que o valor absoluto de ambas as funções decai ao longo dos $lags$

seguindo o formato de uma senoide amortecida (tanto a parte sazonal quanto a não sazonal), o que é aderente à utilização de ambos os componentes, de médias móveis e autorregressivo.

Dessa maneira, o modelo que será ajustado à série D_t será um SARIMA $(p, 0, q) \times (P, 0, Q)_{12}$, o que é equivalente à utilização de um modelo SARIMA $(p, 0, q) \times (P, 1, Q)_{12}$ à série R_t , dado que a série anterior é obtida a partir da diferenciação sazonal simples ($D = 1$) da série de retornos.

Mesmo com a análise visual dos gráficos da fac e da $facp$, fica muito difícil definir quais são os parâmetros p , q , P e Q que mais são aderentes aos dados, já que o comportamento destas funções se apresentou relativamente complexo. Assim, na fase de estimação e seleção, serão testadas diferentes combinações destes parâmetros, com cada um deles variando entre 0 e 2. Conforme mencionado anteriormente, dificilmente é necessário um valor alto para os parâmetros p e q no caso de uma modelagem que combine autorregressão e média móvel.

5.3. Estimação e seleção do modelo

Usando a função “*arima*” da biblioteca “*stats*” do RStudio, foram estimados e avaliados modelos para a série R_t (com o parâmetro d fixado em 0 e D fixado em 1) com cada uma das combinações dos quatro parâmetros, usando os procedimentos especificados no Capítulo 4. Após a estimação de um modelo, também foi calculado seu AIC, utilizando a função “*AIC*” da mesma biblioteca, para avaliação e comparação entre os candidatos. Ao fim das 81 iterações, os modelos foram ranqueados de acordo com o AIC. A Tabela 1 mostra os cinco modelos que obtiveram os melhores desempenhos.

Tabela 1 - AIC dos cinco modelos mais bem classificados

Modelo	p	q	P	Q	AIC
1	2	2	1	2	-941,93
2	2	2	2	2	-940,16
3	2	2	0	2	-940,13
4	2	2	2	1	-940,10
5	2	2	0	1	-939,91

Fonte: Elaborado pelo autor.

Para a seleção do modelo, porém, é necessário ainda avaliar a significância dos coeficientes. Nessa etapa foi utilizada a função “*coefest*” do pacote “*lmtest*” do

RStudio, que retorna uma tabela com o grau de significância de cada um dos parâmetros do modelo. Seguindo a ordem definida pelo AIC, iniciou-se este procedimento com os modelos 1 e 2. Ambos, após a retirada dos coeficientes que não eram significativos ao nível de 5%, apresentaram forte dependência entre os resíduos, o que fez com que eles fossem desconsiderados.

Com isso, o terceiro modelo mais bem classificado, SARIMA $(2,0,2) \times (0,1,2)_{12}$, passou a ser analisado. O teste de seus coeficientes retornou que apenas o segundo coeficiente sazonal de médias móveis não era significativo ao nível de 5%, com um p-valor de 0,14, aproximadamente, conforme a Tabela 2.

Tabela 2 - Teste de coeficientes do modelo 3

Coeficiente	Estimativa	Desvio padrão	z	p-valor ($> z $)
ar1	-0,881	0,138	-6,381	<1e-03
ar2	-0,671	0,119	-5,648	<1e-03
ma1	0,412	0,168	2,453	0,014
ma2	0,373	0,141	2,642	0,008
sma1	-0,668	0,097	-6,901	<1e-03
sma2	-0,146	0,099	-1,476	0,140

Fonte: Elaborado pelo autor.

Excluindo este coeficiente do modelo, temos como resultado um SARIMA $(2,0,2) \times (0,1,1)_{12}$, semelhante ao quinto melhor modelo conforme mostra a Tabela 1. Realizando o teste de significância novamente, vemos que todos estes coeficientes são significativos ao nível de 5%, conforme apresentado na Tabela 3.

Tabela 3 - Teste de coeficientes do modelo 3 reduzido

Coeficiente	Estimativa	Desvio padrão	z	p-valor ($> z $)
ar1	-0,941	0,104	-9,042	<1e-03
ar2	-0,736	0,101	-7,281	<1e-03
ma1	0,503	0,139	3,629	<1e-03
ma2	0,417	0,135	3,090	0,002
sma1	-0,781	0,063	-12,346	<1e-03

Fonte: Elaborado pelo autor.

O AIC do novo modelo é, evidentemente, o mesmo do modelo 5: -939,91. Assim, o único candidato que ainda poderia mostrar-se superior a ele era o modelo 4, um SARIMA $(2,0,2) \times (2,1,1)_{12}$. Realizando o mesmo procedimento, a função “coeftest” retornou que ambos os coeficientes sazonais de autorregressão não eram significativos, o que resultou novamente no modelo 5. Assim, o modelo selecionado para avançar à fase do diagnóstico foi este. A expressão do modelo 5 pode ser vista na equação (5.1):

$$D_t = -0,941D_{t-1} - 0,736D_{t-2} + a_t + 0,503a_{t-1} + 0,417a_{t-2} - 0,781a_{t-12}. \quad (5.1)$$

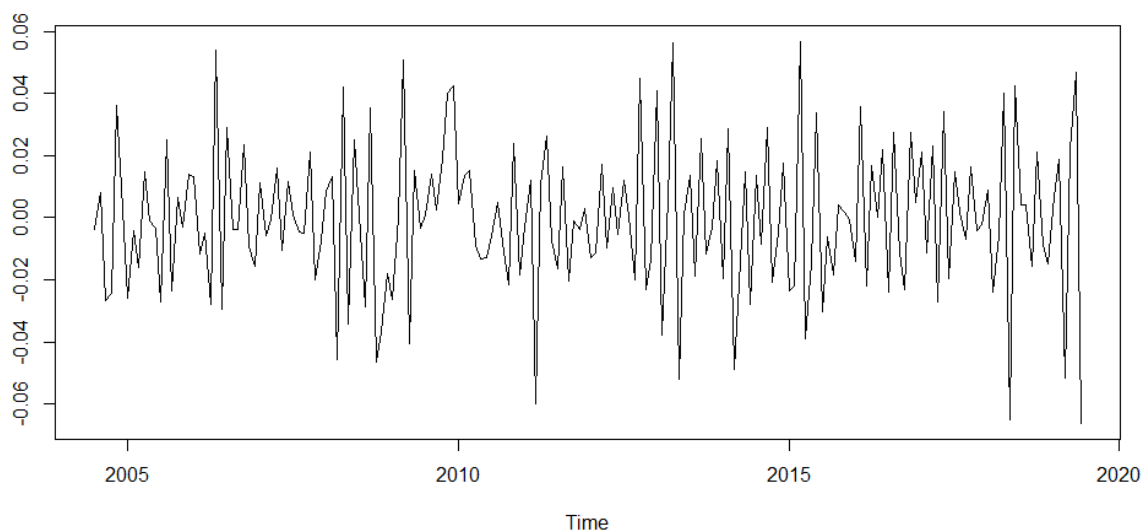
Relembrando que a série D_t é obtida a partir da diferenciação sazonal da série de retornos mensais, R_t :

$$D_t = \Delta^{12}R_t = R_t - R_{t-12}. \quad (5.2)$$

5.4. Diagnóstico do modelo

Com o modelo SARIMA aderente à série de retornos do IBC-Br devidamente identificado e estimado, resta a fase de diagnóstico. Nesta etapa, conforme apresentado no Capítulo 4, os resíduos a_t do modelo em questão serão testados de modo a verificar se eles se comportam como ruído branco. Os valores de a_t retornados pelo modelo estão ilustrados na Figura 22.

Figura 22 - Resíduos do modelo selecionado



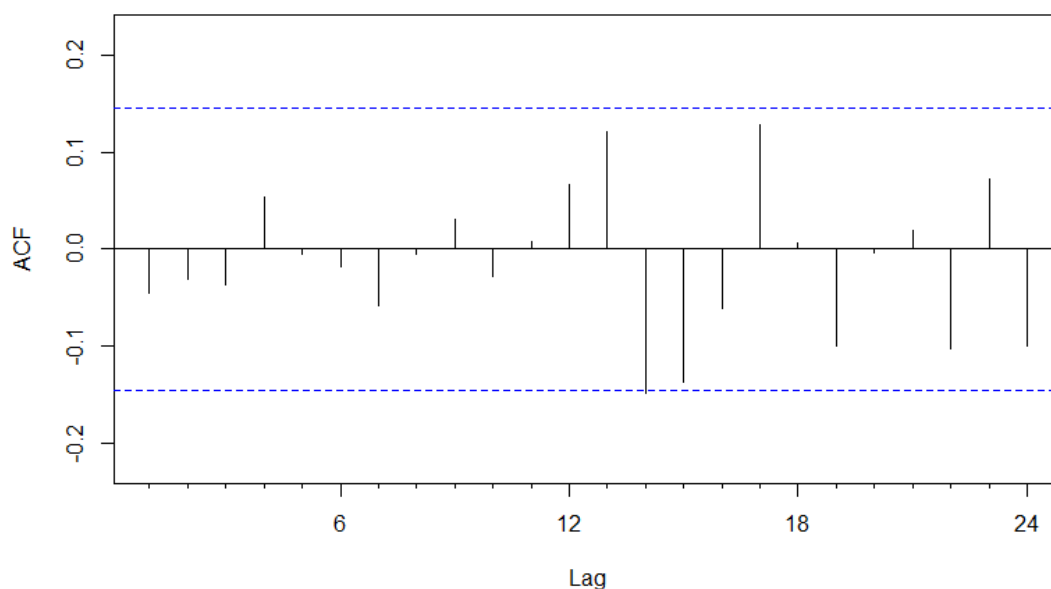
Fonte: Elaborado pelo autor.

Podemos notar, a partir da Figura 22, que os resíduos parecem ser bem distribuídos ao longo dos períodos, com uma média constante e próxima de zero, o que é aderente ao comportamento de um ruído branco. Calculando a média dos a_t , verifica-se que o valor absoluto dela é realmente bem baixo, de apenas 0,002.

A próxima propriedade a ser testada é a de independência entre os resíduos. Para tal, as funções de autocorrelação e autocorrelação parcial da série dos resíduos foram calculadas e dispostas em gráficos, os quais estão exibidos nas Figuras 23 e 24. É possível perceber que os valores de fac e $facp$ são consideravelmente menores que os valores destas funções para a série D_t , que tinha um número elevado de fac e $facp$ significativos, muitos deles com valores absolutos acima de 0,2. Além disso, o comportamento dos a_t neste quesito assemelha-se muito ao comportamento de um

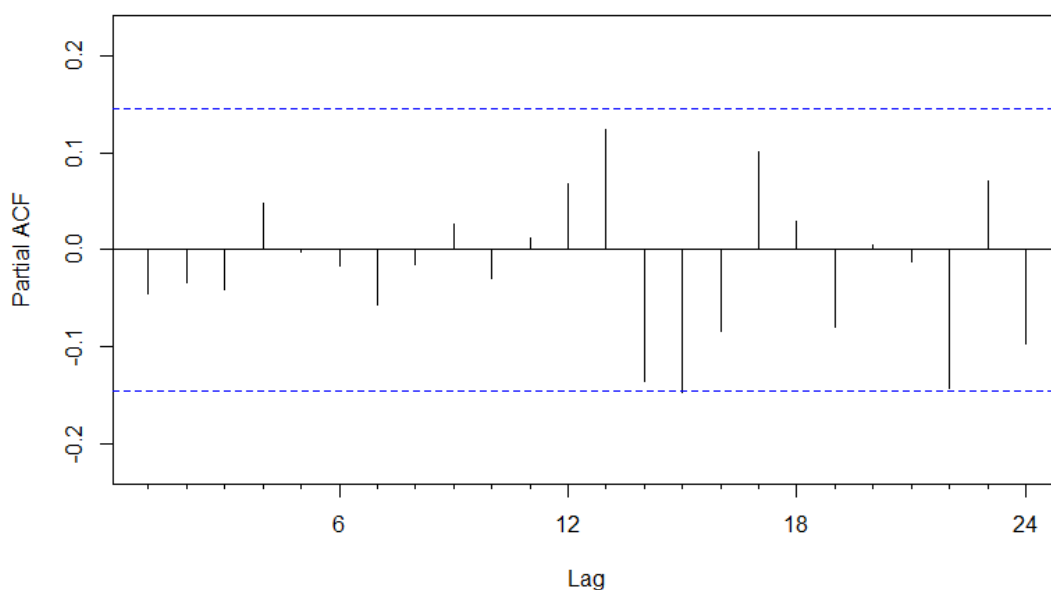
ruído branco, com apenas um *lag* mostrando valores significativos para a *fac* (*lag* 14) e para a *facp* (*lag* 15), e mesmo estes assumindo valores muito próximos ao limite de significância de $\pm 0,15$. De qualquer maneira, para chegar a uma conclusão mais assertiva acerca da independência dos resíduos, é importante realizar outro teste para esta propriedade.

Figura 23 - *Fac* dos resíduos do modelo



Fonte: Elaborado pelo autor.

Figura 24 - *Facp* dos resíduos do modelo

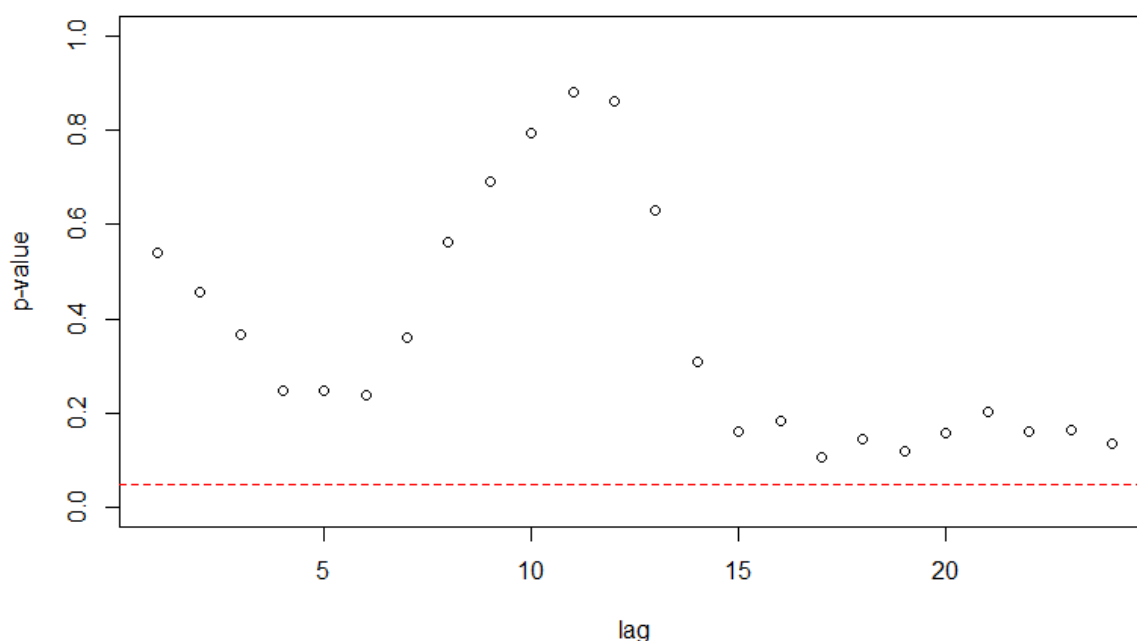


Fonte: Elaborado pelo autor.

Conforme apresentado no Capítulo 4, o teste estatístico proposto para verificar a independência dos resíduos é o do Ljung-Box. Utilizando a função “Box.test” da biblioteca “stats” no RStudio, o teste em questão foi aplicado aos resíduos do modelo,

retornando um p-valor de 0,14. Assim, não podemos rejeitar a hipótese nula de independência ao nível de 5%, e os resíduos podem ser considerados independentes e identicamente distribuídos. Visualizando os resultados do teste de Ljung-Box graficamente, fica claro que não há evidência para rejeitar H_0 para nenhum dos lags entre os a_t , conforme mostra a Figura 25, com todos os p-valores acima da linha de significância traçada em 0,05.

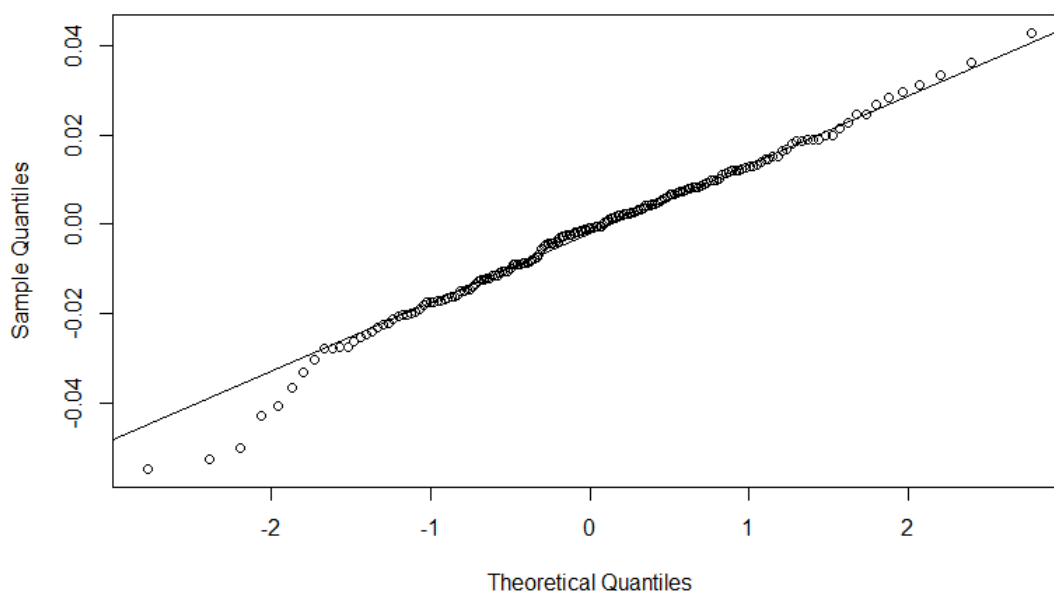
Figura 25 - Representação gráfica do teste de Ljung-Box



Fonte: Elaborado pelo autor.

A última propriedade que deve ser testada para que seja verificado se os resíduos se comportam como ruído branco é a normalidade. Para tal, primeiramente foi construído o gráfico QQ dos a_t , apresentado na Figura 26. Observa-se que a maior parte dos pontos é aderente à reta normal, com exceção dos poucos valores localizados no extremo inferior dela.

Figura 26 - Gráfico QQ dos resíduos



Fonte: Elaborado pelo autor.

Ainda testando a distribuição dos resíduos, foi realizado o teste estatístico de Shapiro-Wilk sobre eles, através da função “shapiro.test” do pacote “stats” do RStudio. O valor da estatística W retornado foi de 0,988, equivalente a um p-valor de 0,11. Dessa maneira, não se pode rejeitar a hipótese nula de normalidade a um nível de 5%, e assume-se que os resíduos seguem uma distribuição Gaussiana.

Com os resultados favoráveis obtidos nos testes realizados sobre os resíduos, é possível assumir que eles seguem o comportamento de ruído branco. Assim, podemos concluir que o modelo estimado é adequado para a modelagem da série do IBC-Br, e será utilizado para a previsão e geração dos gráficos de controle.

5.5. Determinação dos limites de controle

Com o modelo para o IBC-Br devidamente validado, é possível avançar para a fase de determinação dos limites de controle do gráfico EWMA. Conforme especificado na Seção 4.5, o valor de λ que será utilizado é de 0,6, e o valor alvo do ARL do processo será de 36 (o que equivale a uma probabilidade de erro do tipo I de $1/36$, ou 2,8%), indicando uma observação fora de controle a cada 36 meses, em média. O valor central do gráfico de controle foi definido em 0, dado que a média dos resíduos do modelo é muito próxima deste valor. Além dessas, foram definidos os parâmetros a seguir para a realização do algoritmo:

- $N = 10000$, de modo a realizar um número elevado de simulações;

- $L_{ini} = 0$;
- $step = 0,01$.

A partir disso, o algoritmo para cálculo dos limites de controle pôde ser iniciado, com a simulação dos resíduos $a_t \sim N(0, \sigma^2)$, com $\sigma^2 = 0,0002742479$. Para cada iteração, foi estimada uma série X_t de 1000 observações com os coeficientes do modelo estimado para D_t . Estas séries foram geradas a partir da função “sarima.sim”, da biblioteca “astsa” do RStudio.

Conforme o algoritmo explicitado na Seção 4.5, os X_t são utilizados para cálculo dos Z_t , que por sua vez são comparados aos limites de controle para a obtenção do *run length* do processo. O primeiro valor de L que apresentar um ARL acima de 36 será selecionado para utilização no gráfico de controle. Após dezenas de iterações, o algoritmo retornou um valor para L_{fim} de 1,73, o que indica que os limites de controle do gráfico para Z_t serão fixados em:

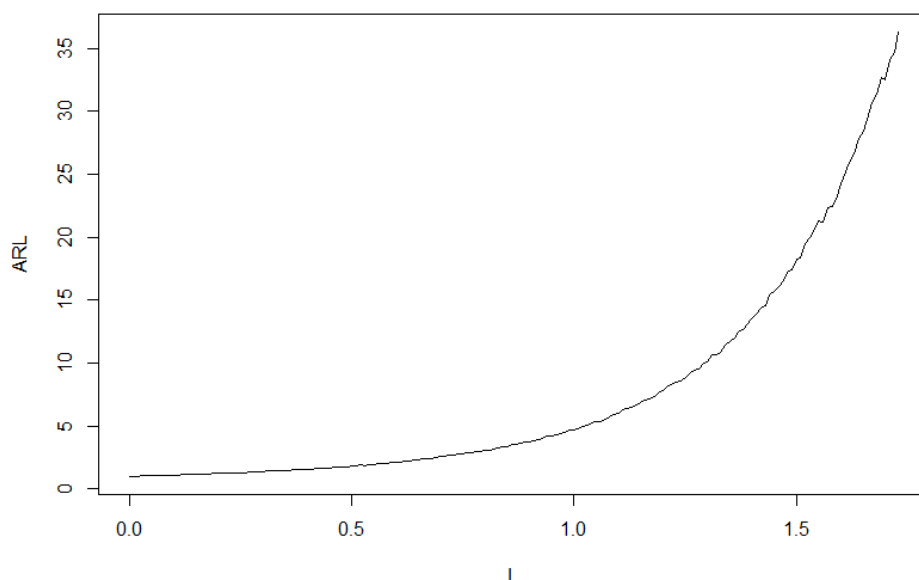
$$LIC = 0 + 1,73\sqrt{0,0002742479} = 0,02864955$$

$$LC = 0$$

$$LSC = 0 - 1,73\sqrt{0,0002742479} = -0,02864955$$

O ARL encontrado para este valor de L foi de 36,28, o que indica que, em média, o gráfico EWMA identificará uma observação fora de controle a cada 36 ou 37 meses, ou aproximadamente três anos. O gráfico exibido na Figura 27 mostra como o ARL aumentou à medida que o valor de L era incrementado por *step*.

Figura 27 - ARL em função de L



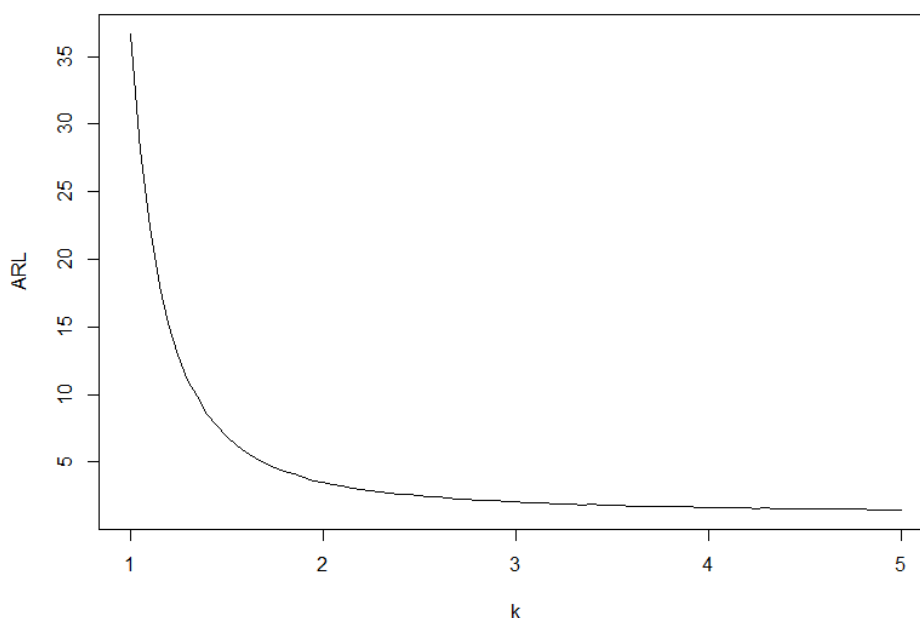
Fonte: Elaborado pelo autor.

5.6. Obtenção dos ARL fora de controle

Após a definição do parâmetro L , que determina os limites de controle do gráfico EWMA, passa a ser possível analisar o desempenho da ferramenta ao ser aplicada em um processo fora de controle, através da avaliação do ARL_1 . Para tal, conforme especificado na Seção 4.6, deve ser executado um algoritmo semelhante ao da etapa anterior, mas com valores distintos para o desvio padrão dos resíduos simulados, e com o parâmetro k variando ao invés de L .

Dessa maneira, os resíduos foram gerados a partir da distribuição $a_t \sim N(0, k\sigma^2)$. O parâmetro k foi variado de 1 a 5, com incrementos de 0,05, para analisar a sensibilidade do gráfico em identificar processos fora de controle. A primeira iteração, portanto, é igual à situação final do algoritmo da seção anterior. Assim, espera-se que o ARL decaia rapidamente de 36,28 para valores menores. Os resultados obtidos podem ser visualizados na Figura 28.

Figura 28 - ARL em função de k



Fonte: Elaborado pelo autor.

A partir da análise da Figura 28, é possível concluir que o gráfico de controle mostrou-se eficiente em detectar variações na série do IBC-Br. Apenas com o incremento de k de um (situação normal) para dois, o ARL médio foi de 36,28 para 3,48, indicando uma redução de 90,4%.

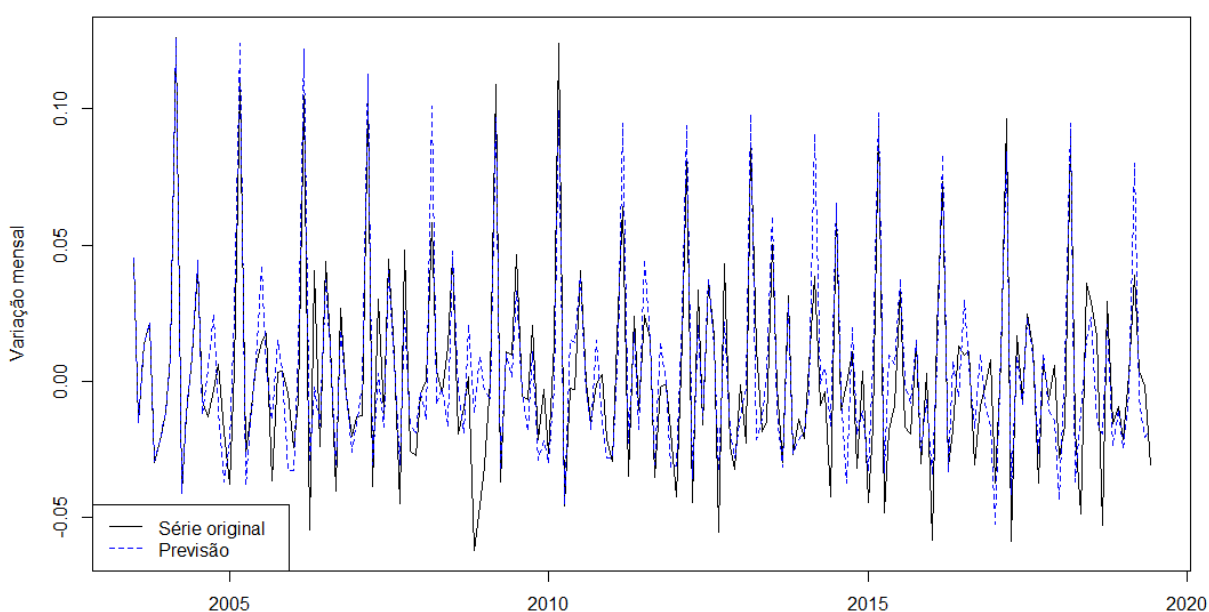
5.7. Avaliação dos resultados

Conforme relatado no Capítulo 4, os principais objetivos de realizar a modelagem do IBC-Br por séries temporais, no presente trabalho, são dois: realizar previsões dos valores futuros da série, de modo a ter um direcionamento de qual a tendência do nível de atividade da economia brasileira no curto e médio prazo; e realizar seu controle estatístico através de gráficos de controle, identificando os momentos nos quais a atividade está abaixo ou acima do padrão.

5.7.1. Previsões dos valores futuros

Assim, para avaliar os resultados, primeiramente é necessário avaliar a eficácia do modelo SARIMA $(2,0,2) \times (0,1,1)_{12}$ ajustado à série de retornos R_t em utilizar os valores passados da série de maneira a prever seus valores futuros. Para tal, a primeira análise realizada foi a comparação da série de treinamento, contendo os dados de 2003 a 2019, com os valores calculados pelos parâmetros do modelo. Os resultados estão apresentados na Figura 29. A série original exibida na figura é a série de variação mensal, R_t , e a linha azul pontilhada representa as previsões para essa variável.

Figura 29 - Série de treinamento vs. valores ajustados pelo modelo



Fonte: Elaborado pelo autor.

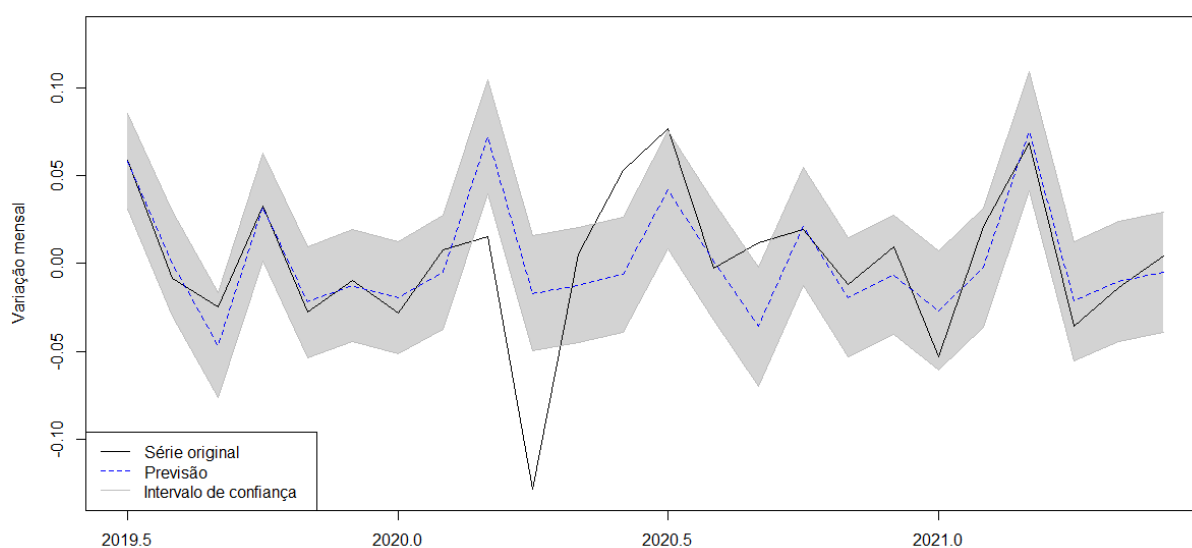
Podemos notar que os dados ajustados pelo modelo parecem ser bem aderentes à série, inclusive com os picos e vales causados pela sazonalidade

identificados na maioria dos períodos. Além disso, é possível avaliar o desempenho do modelo quantitativamente, através do RMSE. Utilizando a função “rmserr” da biblioteca “pracma” do RStudio, o valor retornado para o RMSE entre os valores ajustados e observados foi de 0,016. Dado que a amplitude dos valores observados da série é de 0,188, o desempenho do modelo na série de treinamento parece satisfatório, com o RMSE assumindo um valor menor que 10% da amplitude da série.

Porém, para avaliar o desempenho do modelo em prever as variações mensais futuras do IBC-Br – que é o que realmente estamos interessados –, é preciso estudar suas previsões para a série de teste, que não foi utilizada para o ajuste do modelo. É esperado que o desempenho seja inferior ao observado na série de treinamento, como já explicado no Capítulo 4, mas que ainda assim o modelo tenha um bom poder preditivo. Outro fator que dificulta a previsão é o horizonte relativamente longo, de dois anos completos.

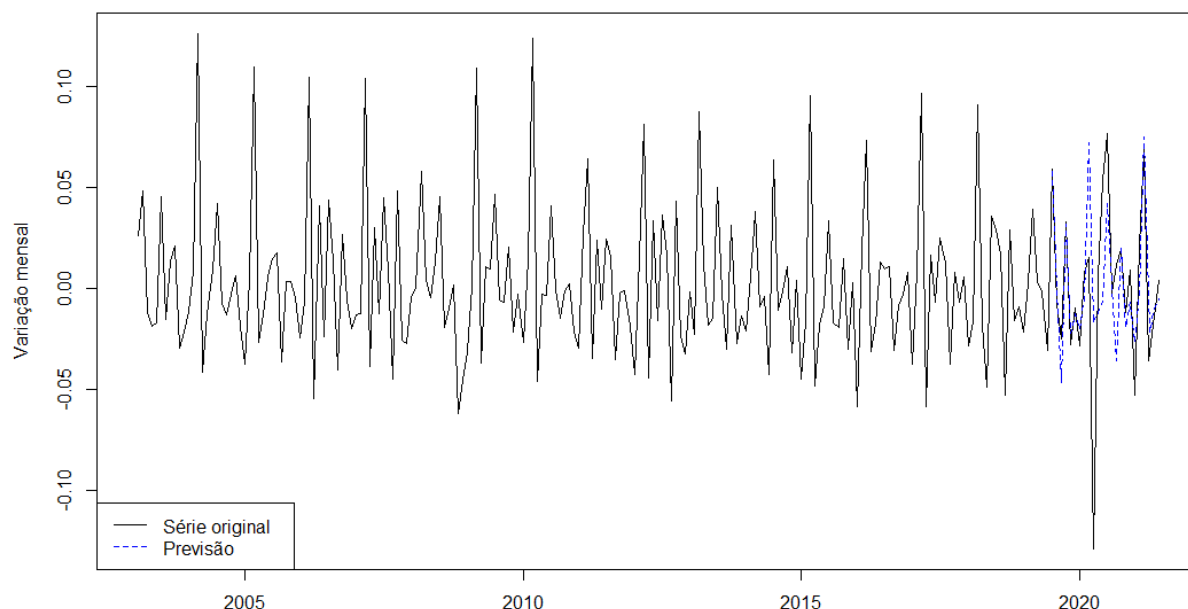
Utilizando a função “predict” do pacote “stats”, foram feitas as previsões, baseadas no modelo ajustado, das 24 observações da série de treinamento, de julho de 2019 a junho de 2021. Os valores da série R_t , suas previsões e os intervalos de confiança de 90% estão exibidos na Figura 30. A Figura 31 traz a série completa em conjunto com as previsões realizadas.

Figura 30 - Série de teste vs. previsões feitas pelo modelo com intervalo de confiança de 90%



Fonte: Elaborado pelo autor.

Figura 31 - Série completa vs. previsões feitas pelo modelo



Fonte: Elaborado pelo autor.

Como podemos observar, a maioria das previsões ficou dentro dos intervalos calculados. Das 24 observações previstas, 19 ficaram dentro dos intervalos de confiança, indicando uma assertividade de aproximadamente 79%. Mesmo com esse número ficando abaixo de 90%, o desempenho do modelo pode ser avaliado como satisfatório – ainda mais ao se considerar que foram feitas previsões para um horizonte relativamente longo –, dado que ele poderia ser ainda melhor não fosse um evento inesperado que aconteceu no primeiro semestre de 2020: a pandemia de Covid-19.

O gráfico da série de teste vs. previsões do modelo mostra que elas foram muito assertivas ao longo do segundo semestre de 2019 e os primeiros meses do ano seguinte. A partir de março de 2020, entretanto, o desempenho do modelo teve uma queda significativa, o que pode ser explicado pela chegada da Covid-19 no Brasil e toda a incerteza relacionada a ela, que acarretou o fechamento dos comércios no país inteiro e a consequente redução brusca do nível de atividade econômica. Evidentemente, o modelo não conseguiu prever essa quebra, o que fez com que diversas das observações deste período ficassem fora dos intervalos de confiança de 95%. Quatro (março, abril, junho e julho de 2020) das cinco observações que ficaram fora dos intervalos de confiança aconteceram nos primeiros meses da pandemia, quando a incerteza acerca da situação era elevada. A outra foi em setembro do mesmo ano, quando a situação estava voltando a se estabilizar. Com a efetiva

estabilização da atividade no final de 2020, é visível que o modelo volta a atingir uma boa performance, mesmo com o horizonte de previsão longo, com todas as observações deste ano dentro dos intervalos de confiança. Excluindo os seis meses de maior incerteza relacionada à Covid-19 (março a agosto de 2020), 17 de 18 previsões ficaram dentro dos intervalos de confiança, indicando um nível de acerto de 94%, acima do nível de 90% determinado pelos intervalos.

As previsões para a série de teste obtiveram um RMSE de 0,033, indicando um desempenho pior que os valores ajustados para a série de treinamento. Entretanto, essa diferença pode ser atribuída a diversos motivos: a maior amplitude da série de teste do que da série de treinamento (0,205 contra 0,188), mesmo tendo menos observações, o que indica alta variabilidade; a quebra do padrão encontrado anteriormente por conta da pandemia; e a natural piora no desempenho entre as séries de treinamento e teste, dado que a segunda não é utilizada na modelagem dos dados.

Ainda assim, o modelo mostra-se bastante eficiente para a realização das previsões, com um RMSE de apenas 16% da amplitude da série e boa acuracidade dos intervalos de confiança, mesmo com as observações tendo sido afetadas pela Covid-19. Por essas razões, o modelo proposto pode ser considerado um bom estimador para os valores futuros da série de variações mensais do IBC-Br, atuando como uma ferramenta útil para a tomada de decisão no contexto de ativos e investimentos que podem ser afetados pelo nível de atividade econômica brasileira, conforme discutido na Introdução.

5.7.2. Gráfico de controle

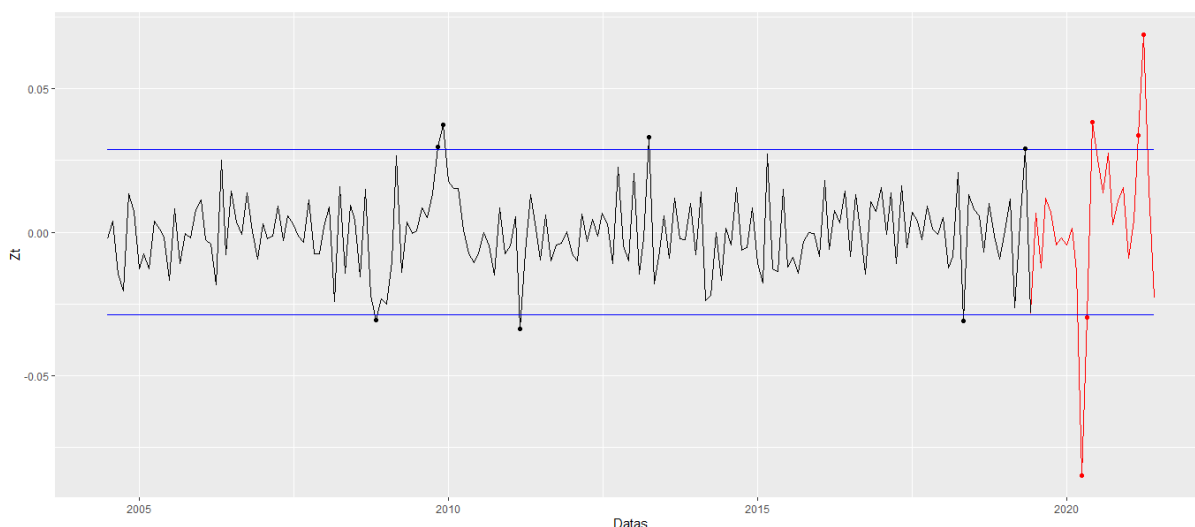
Em seguida, é necessário analisar a segunda parte do modelo proposto, que é a do gráfico de controle EWMA aplicado sobre a série estacionária D_t , à qual foi ajustada o modelo SARIMA. Para isso, é preciso calcular a série de estatísticas Z_t a partir dos dados reais da série utilizada, para o período entre julho de 2004 e junho de 2021, correspondentes aos dados de treinamento e de teste. Já os limites de controle foram definidos na Seção 5.5, apresentando os valores de aproximadamente -0,029 para o LIC e 0,029 para o LSC (com a LC em 0).

Como todos os parâmetros do modelo já foram estimados e estamos apenas visualizando os resultados do gráfico de controle, iremos utilizar também os dados de

teste, de maneira a avaliá-los de forma mais aprofundada. Outro ponto importante é avaliar o comportamento do gráfico de controle nos meses de 2020 durante os quais a pandemia de Covid-19 mais afetou a atividade da economia brasileira, entre março e agosto. Espera-se que o gráfico EWMA indique, nos meses iniciais da pandemia, que a estatística Z_t estava fora de controle neste período, apresentando valores abaixo do esperado pelo modelo.

A Figura 32 apresenta os resultados obtidos com a aplicação do gráfico EWMA sobre a estatística Z_t . A série na cor preta corresponde aos dados de treinamento, e a série na cor vermelha aos dados de teste. As linhas azuis indicam os limites de controle.

Figura 32 - Gráfico de controle EWMA da estatística Z_t



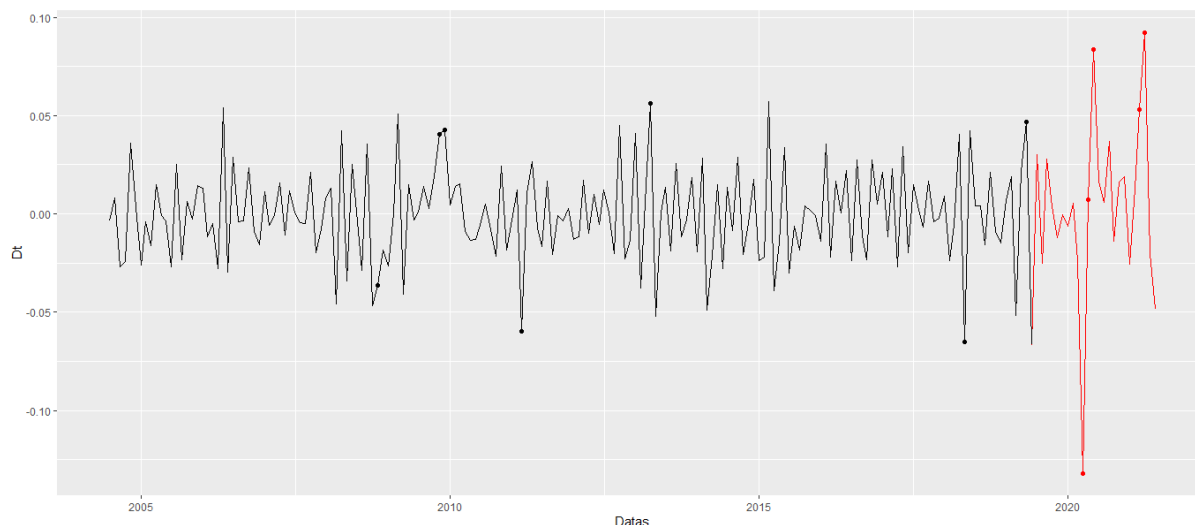
Fonte: Elaborado pelo autor.

Como é possível observar, o monitoramento da série do IBC-Br pelo gráfico de controle EWMA identificou 12 períodos fora de controle. Destes, cinco foram identificados como períodos durante os quais a atividade econômica brasileira desempenhou abaixo do esperado: novembro de 2008, março de 2011, maio de 2018 e abril e maio de 2020. Os demais sete foram sinalizados como períodos nos quais a atividade econômica do país obteve desempenho superior ao esperado pelo modelo: novembro e dezembro de 2009, abril de 2013, maio de 2019, junho de 2020 e março e abril de 2021.

Evidentemente, os pontos que foram identificados pelo gráfico EWMA são os que apresentaram os maiores valores absolutos para a estatística Z_t . Porém, ao se destacar estes 12 meses sobre a série que foi modelada, D_t , podemos notar que a

identificação de pontos fora de controle não é tão simples como parece. A Figura 33 apresenta esta visualização.

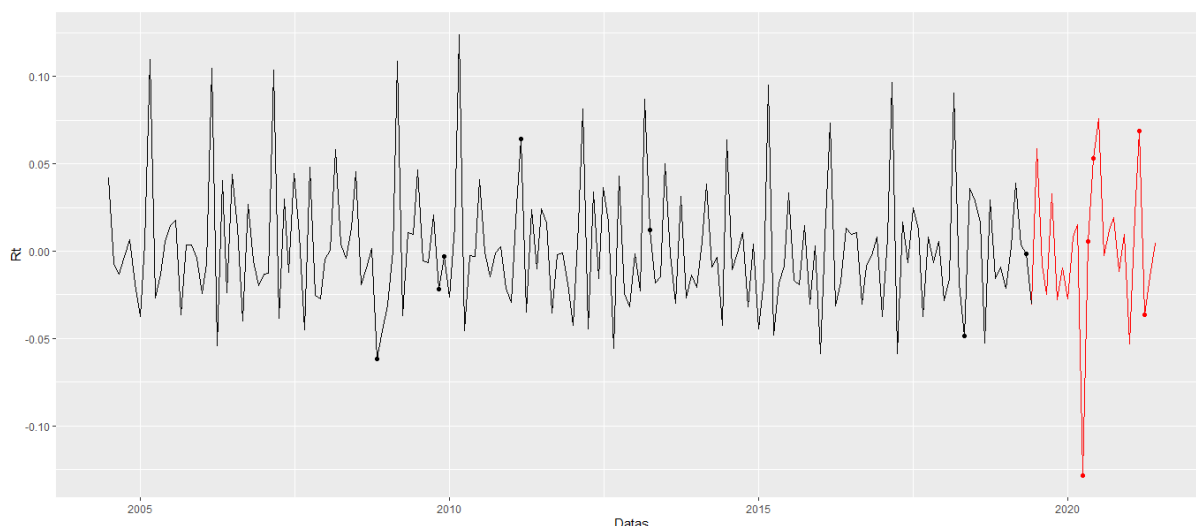
Figura 33 – Série estacionária D_t com destaque para os pontos fora de controle



Fonte: Elaborado pelo autor.

Os períodos destacados como fora de controle não são necessariamente aqueles cujos valores absolutos são os mais elevados, o que indica que a associação do modelo SARIMA com o cálculo da estatística Z_t permitiu que fossem identificados períodos fora de controle que, à primeira vista, parecem estar dentro do esperado. Assim, é possível observar na prática a vantagem do gráfico de controle EWMA sobre o de Shewhart em detectar variações menores da média do processo, ao levar em consideração não apenas a observação atual, mas também o histórico do processo. Se o segundo gráfico tivesse sido empregado na análise, seriam identificados como fora de controle apenas aqueles valores de D_t acima ou abaixo de determinado patamar.

Foi também gerada uma terceira visualização para os períodos identificados como fora de controle, exibida na Figura 34. Nela, temos destacados os pontos sinalizados pelo gráfico EWMA sobre a série original de retornos, R_t , que possui elevadas variações causadas pela sazonalidade da série.

Figura 34 - Série de retornos R_t com destaque para os pontos fora de controle

Fonte: Elaborado pelo autor.

A partir da análise deste gráfico, fica evidente a importância de se levar em consideração a sazonalidade da série ao fazer sua modelagem. Caso a série R_t não houvesse sido diferenciada sazonalmente, resultando na série D_t , os gráficos de controle teriam tido dificuldade em identificar os pontos que realmente estavam fora do comportamento esperado da série, por conta da alta variabilidade que acontece naturalmente ao longo do ano. Podemos notar que diversos dos pontos sinalizados pelo gráfico estão longe dos valores mais extremos, valores estes decorrentes da sazonalidade, e não de causas especiais que realmente colocariam o processo fora de controle.

Retornando aos pontos que foram sinalizados pelo gráfico EWMA como fora de controle, é importante analisá-los para relacionar com possíveis fatores externos que ocorreram nos períodos correspondentes. Com isso, será possível verificar se o gráfico de controle conseguiu identificar períodos nos quais alguma causa especial poderia ter feito com que a atividade econômica brasileira ficasse fora de controle.

Iniciando a análise pelas observações cuja estatística Z_t ficou abaixo do LIC, foram identificados eventos atípicos em todos os cinco períodos, que poderiam ter atuado como causas especiais para a redução da atividade econômica no Brasil. Segue a relação dos períodos com os acontecimentos externos:

- Novembro de 2008: a crise financeira dos *subprimes*, que se iniciou nos EUA em 2007 e se difundiu pelo mundo todo nos anos seguintes, teve seu auge em setembro de 2008, com a falência de diversas instituições financeiras do país, entre elas o banco de investimentos Lehman

Brothers. Assim, a queda na atividade econômica no Brasil, dois meses depois, pode indicar os primeiros efeitos da crise sentidos no país.

- Março de 2011: no início deste mesmo mês, houve o terremoto e o tsunami no Japão, que causou grandes destruições ao país e impactou fortemente sua economia. À época, o Japão era o quinto maior destino das exportações brasileiras, e esse desastre pode ter impactado, mesmo que indiretamente, a economia do Brasil.
- Maio de 2018: no dia 21 deste mês, iniciou-se a greve dos caminhoneiros, que perdurou até o dia 30 de maio. A paralisação dos caminhoneiros e os bloqueios realizados nas estradas causaram diversos impactos na atividade do país, como indisponibilidade de alimentos e combustíveis, suspensão de aulas e eventos, cancelamento de voos, entre outros. Assim, era esperado que este período fosse sinalizado pelo gráfico de controle.
- Abril e maio de 2020: por conta da chegada e da disseminação do novo coronavírus ao Brasil no início de 2020, no final de março deste mesmo ano foram decretadas, em todos os estados do país, diversas medidas que visavam o fechamento dos comércios e de outras atividades consideradas não essenciais. Com isso, a economia do Brasil sofreu um grande baque, principalmente nos primeiros meses que foram totalmente impactados por essas medidas.

Com relação aos períodos de alta no IBC-Br, notou-se um comportamento distinto aos de baixa, sem ter ocorrido a identificação de possíveis causas especiais para a maior parte deles. Podemos destacar, aqui, o mês de junho de 2020. Apesar do Brasil ainda estar no auge das restrições de atividade por conta da Covid-19, os meses após o choque inicial da interrupção das atividades mostraram importante recuperação econômica, com o pagamento do auxílio emergencial e o aprimoramento de plataformas de comércio online, entre outros fatores.

Para os demais períodos de alta do IBC-Br, não foram identificados eventos atípicos. Percebe-se, entretanto, que três deles (novembro de 2009, maio de 2019 e abril de 2021) coincidem com períodos um ano após períodos de baixa. Assim, podemos concluir que a identificação destes valores como pontos fora de controle é

um indicativo que a economia brasileira conseguiu se recuperar de maneira positiva depois de períodos de redução na atividade econômica.

Portanto, os resultados do gráfico de controle mostram que ele foi eficiente em identificar períodos nos quais a atividade econômica brasileira evoluiu de maneira atípica, dado que diversas das observações sinalizadas foram em períodos nos quais foi encontrada uma possível causa especial. Assim, podemos concluir que o gráfico EWMA aplicado ao IBC-Br pode ser uma ferramenta interessante a ser utilizada para o monitoramento da economia brasileira e, conseqüentemente, para a gestão de ativos financeiros.

6. CONCLUSÃO

Este capítulo final trará uma breve recapitulação do desenvolvimento deste trabalho de formatura, destacando os resultados alcançados e avaliando se o objetivo proposto na Introdução foi atingido. Serão também apresentadas algumas sugestões de aprofundamento deste estudo, assim como propostas de estudos futuros que auxiliariam no cumprimento do objetivo. Por fim, serão destacadas as principais dificuldades enfrentadas pelo autor ao longo do trabalho.

Ao longo dos estágios realizados durante a graduação em dois fundos de investimentos, o autor deste trabalho observou a importância do acompanhamento de indicadores da economia brasileira, dado que ela afeta grande parte dos ativos financeiros, e notou a necessidade de um algoritmo quantitativo que auxiliasse os gestores na tomada de decisão, com projeções para um horizonte de tempo mais distante. A partir disso, o autor se propôs a desenvolver um modelo de previsão e monitoramento do IBC-Br, um índice mensal que busca medir o nível de atividade econômica do Brasil, através dos conceitos estatísticos de séries temporais e de CEP.

Assim, antes de desenvolver uma metodologia para o modelo em questão, era necessário aprofundar-se no estudo destes dois temas, o que culminou na revisão bibliográfica detalhada ao longo dos Capítulos 2 e 3. Após uma análise inicial sobre os temas, decidiu-se que a metodologia seria baseada em modelos paramétricos de séries temporais da classe SARIMA e em gráficos de controle. No capítulo seguinte, foram definidos os métodos que seriam utilizados para o desenvolvimento do algoritmo, entre eles: EMV e EMQ para a estimação do modelo paramétrico, AIC para a comparação dos modelos candidatos, RMSE para a avaliação das previsões, gráfico de controle EWMA para o monitoramento do processo e um algoritmo iterativo para a definição dos limites de controle.

Em seguida, a metodologia foi aplicada à série do IBC-Br, que foi dividida em dois conjuntos: um para a estimação do modelo, chamado de treinamento, com dados de 2003 a 2019; e outro para a avaliação dele, chamado de teste, com dados de 2019 a 2021. Inicialmente, os dados foram manipulados através da aplicação de retornos e diferenciação de forma a obter-se uma série estacionária à qual um modelo paramétrico pudesse ser ajustado. Por fim, chegou-se a um modelo $SARIMA(2,0,2) \times (0,1,1)_{12}$, aprovado em todas as etapas do diagnóstico.

Com relação ao monitoramento do processo, inicialmente foram determinados os limites de controle do gráfico EWMA, a partir do ARL_{alvo} de 36, do λ de 0,6 e do método iterativo. Com isso, chegou a um limite de controle de $\pm 1,73$ o desvio padrão dos resíduos do modelo. Além disso, as simulações realizadas com o aumento do desvio padrão dos resíduos mostraram que o gráfico em questão era eficiente para identificar processos fora de controle.

O modelo de previsão desenvolvido foi, então, avaliado para que fosse determinado se ele era eficaz em cumprir seu propósito. Os valores previstos para a série de teste tiveram boa aderência aos dados reais, mesmo com a alta instabilidade econômica que o país atravessou nesse período, com a pandemia de Covid-19. Desconsiderando os seis meses iniciais da pandemia, quando os impactos econômicos foram mais intensos no Brasil – o que, evidentemente, não foi previsto pelo modelo –, 94% das previsões ficaram dentro do intervalo de confiança de 90%, o que evidencia sua boa performance.

Avaliou-se, em seguida, o desempenho do gráfico de controle EWMA, que identificou 12 observações nas quais o IBC-Br estava fora de controle, das quais cinco aconteceram após o início da pandemia de Covid-19 – o que evidencia que o gráfico identificou o comportamento fora do padrão da série neste período. Nos períodos sinalizados pelo gráfico, em grande parte deles foi notado um acontecimento externo que poderia ter impactado a economia brasileira, especialmente nos de baixa, um sinal de que a ferramenta identificou de forma eficaz períodos nos quais a atividade econômica brasileira esteve fora de controle.

Com base nos resultados obtidos tanto pelo modelo de previsão quanto pelo gráfico de controle, pode-se concluir que a ferramenta estatística proposta cumpre seu objetivo. O fato de que o IBC-Br ficou muito próximo das previsões do modelo, mesmo para um horizonte de mais de um ano, corrobora isto. Assim, espera-se que, se associado a outras ferramentas, qualitativas e quantitativas, por gestores de portfólio, o modelo desenvolvido possa trazer resultados interessantes, servindo como um indicativo confiável para o andamento futuro da economia brasileira e para seu monitoramento no presente.

Apesar do ótimo desempenho apresentado pelo modelo, há ainda a possibilidade aperfeiçoá-lo através de outras funcionalidades. No caso de previsões para o curto prazo, pode-se adicionar variáveis exógenas ao modelo, com a Pesquisa Industrial Mensal e a Pesquisa Mensal do Comércio, que são liberadas antes do IBC-Br e

poderiam potencializar a capacidade de previsão do modelo ao serem combinadas com a modelagem SARIMA. Ainda visando a melhoria das previsões, a adição de variáveis binárias identificando períodos *outliers* poderia ser benéfica, de modo que as previsões futuras não seriam muito afetadas por quebras inesperadas na variável analisada.

A parte de monitoramento do processo poderia ser aprofundada através da análise de sensibilidade dos parâmetros do gráfico EWMA. Uma sugestão seria avaliar diferentes valores de λ e o desempenho do gráfico de controle para cada um deles, de forma a encontrar a ponderação ótima entre a observação atual e o histórico da série. Outra sugestão seria testar como o gráfico de controle reage à variação da média do processo para a avaliação do ARL fora de controle, ao invés de modificar sua variância, como foi feito no presente trabalho.

Fora as melhorias que podem ser aplicadas ao próprio modelo já desenvolvido, existe a possibilidade do desenvolvimento de estudos futuros que auxiliem no objetivo de monitoramento da economia brasileira. Aqui, sugere-se a aplicação da metodologia proposta a outras variáveis que refletem o comportamento da economia do Brasil, como o PIB, o IPCA e o Ibovespa. Nesse último caso, pode existir uma complexidade maior, por ser uma variável diária.

No que se refere às dificuldades encontradas ao longo da elaboração do trabalho, podemos destacar três. Primeiramente, o desenvolvimento de um trabalho acadêmico de grande porte apresentou-se como particularmente desafiador para o aluno, devido à necessidade de leitura de diversas referências bibliográficas e apresentação delas de modo claro e resumido, e à estruturação do texto final, mantendo uma boa coesão entre as ideias do começo ao fim. Outra dificuldade encontrada surgiu ao longo da etapa de identificação e seleção do modelo paramétrico de série temporal. Foram diversas horas gastas simulando diferentes classes de modelos, selecionando os parâmetros que estariam presentes nele, fazendo o diagnóstico de cada um, até que fosse encontrado o melhor modelo aplicável aos dados do IBC-Br. Por fim, a elaboração do algoritmo para a definição dos limites de controle foi outro desafio encontrado pelo autor, dado que, apesar de já ter tido contato com o tema durante a graduação, nunca havia combinado a simulação de uma série temporal a um algoritmo iterativo voltado para tal finalidade.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

BANCO CENTRAL DO BRASIL. Aspectos metodológicos e comparações dos comportamentos do IBC-Br e do PIB. **Relatório de Inflação**, v. 20, n. 1, p. 27–29, 2018.

BOX, George E. P.; JENKINS, Gwilym M. **Time series analysis: forecasting and control**. 1ª ed. San Francisco: Holden-Day, 1970.

BOX, George E. P.; JENKINS, Gwilym M.; REINSEL, Gregory C. **Time series analysis: forecasting and control**. 3ª ed. Englewood Cliffs, NJ: Prentice Hall, 1994. 598 p.

BOX, George E. P.; JENKINS, Gwilym M.; REINSEL, Gregory C.; LJUNG, Greta M. **Time series analysis: forecasting and control**. 5ª ed. Hoboken, NJ: John Wiley & Sons, Inc, 2016. 669 p.

BOX, George E. P.; PIERCE, David A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. **Journal of the American Statistical Association**, v. 65, n. 332, p. 1509–1526, 1970.

COWPERTWAIT, Paul S. P.; METCALFE, Andrew V. **Introductory time series with R**. Dordrecht; New York: Springer, 2009. 254 p.

DICKEY, David A.; FULLER, Wayne A. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. **Journal of the American Statistical Association**, v. 74, n. 366, p. 427, 1979.

HANKE, John E.; WICHERN, Dean W. **Business forecasting**. 9ª ed. Upper Saddle River, NJ: Pearson/Prentice Hall, 2009. 551 p.

KNOTH, Sven; SCHMID, Wolfgang. **Control Charts for Time Series: A Review**. In: LENZ, Hans-Joachim; WILRICH, Peter-Theodor. **Frontiers in Statistical Quality Control 7**. Heidelberg: Physica-Verlag HD, 2004, p. 210–236.

LUCAS, James M.; SACCUCCI, Michael S. Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. **Technometrics**, v. 32, n. 1, p. 1–12, 1990.

MONTGOMERY, Douglas C. **Introduction to statistical quality control**. 7^a ed. Hoboken, NJ: Wiley, 2013. 754 p.

MORETTIN, Pedro A.; TOLOI, Clélia M. C.; **Análise de Séries Temporais**. 2^a ed. São Paulo: Editora Blucher, 2006. 538 p.

ROBERTS, S. W. Control Chart Tests Based on Geometric Moving Averages. **Technometrics**, v. 1, n. 3, p. 239–250, 1959.

SAID, Said E.; DICKEY, David A. Testing for unit roots in autoregressive-moving average models of unknown order. **Biometrika**, v. 71, n. 3, p. 599–607, 1984.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3/4, p. 591, 1965.

YAP, B. W.; SIM, C. H. Comparisons of various types of normality tests. **Journal of Statistical Computation and Simulation**, v. 81, n. 12, p. 2141–2155, 2011.

APÊNDICE A – Código para ajuste do modelo à série do IBC-Br e realização de previsões

```
## Carregando as bibliotecas
```

```
library(pracma)
```

```
library(sidrar)
```

```
library(jsonlite)
```

```
library(dplyr)
```

```
library(quantmod)
```

```
library(forecast)
```

```
library(tseries)
```

```
library(lmtest)
```

```
library(stats)
```

```
library(FitAR)
```

```
## Função para calcular o retorno
```

```
ret <- function(serie) {
```

```
  retorno = (serie - lag(serie)) / lag(serie)
```

```
  return(retorno)
```

```
}
```

```
## Coleta e análise dos dados da série do IBC-Br ----
```

```
## Coletando e manuseando os dados
```

```
ibc =
```

```
fromJSON("http://api.bcb.gov.br/dados/serie/bcdata.sgs.24363/dados?formato=json&  
amp;dataInicial=01/01/2001")
```

```
ibc$data = as.Date(ibc$data, format = "%d/%m/%Y")
```

```
ibc$valor = as.numeric(ibc$valor)
```

```
ibc$var = ret(ibc$valor)
```

```
ibc = ibc[-1,]
```

```
ibc_tudo = ibc[c(1:221),]
```

```
# Os últimos 2 anos de dados (~10%) vão ser usados para testar o modelo
```

```
# Treinamento: jul/2003 até jun/2019
```

```

# Teste: jul/2019 até jun/2021
max = 197
ibc = ibc[c(6:max),]
plot(ibc$valor,type="l")
ibc_nivel = ibc$valor
adf.test(ibc_nivel)
colnames(ibc) = c("datas","nivel","var")

## Criando a série de variação mensal
ibc_var = ts(ibc$var, start = c(2003,7), frequency = 12)
plot(ibc_var, ylab="Variação", xlab="")
plot(decompose(ibc_var))

## Série diferenciada
ibc_des = ibc_var %>% diff(lag=12)
plot(ibc_des,type="l",ylab="R_t - R_t-12",xlab="")

## Escolher entre a série diferenciada ou a série de retornos
ibc_plot = ibc_des

## Teste ADF para ver se a série a ser modelada é estacionária
adf.test(ibc_plot)

plot(decompose(ibc_des))

## Testando a normalidade dos retornos
hist(ibc_plot, breaks = seq(from = -0.16, to = 0.14, by = 0.005))
qqnorm(ibc_plot)
qqline(ibc_plot)
shapiro.test(ibc_plot)

## Verificando a autocorrelação da série
Acf(ibc_plot,lag=24)
Pacf(ibc_plot,lag=24)

```

```
## Modelagem ARIMA ----
```

```
## Teste de todos os modelos candidatos
```

```
for (p in c(0:2)){
  for (q in c(0:2)){
    for (P in c(0:2)){
      for (Q in c(0:2)){
        tryCatch({
          mod_teste = arima(ibc_var, order = c(p,0,q), seasonal = c(P,1,Q))
        }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
        aic = AIC(mod_teste)
        print(c(p,q,P,Q))
        print(aic)
      }
    }
  }
}
```

```
## Modelo final
```

```
novo_modelo <- arima(ibc_des, order=c(2,0,2), seasonal = c(0,0,1),
                     include.mean = F, fixed = c(NA,NA,NA,NA, NA))
coeftest(novo_modelo)
AIC(novo_modelo)
```

```
## Resíduos
```

```
residuos = residuals(novo_modelo)
plot(ibc_des)
plot(residuos, ylab="Resíduos",xlab="")
mean(residuos)
```

```
## Independência dos resíduos
```

```
Acf(residuos)
Pacf(residuos)
```

```

Box.test(residuos, lag = 24, type = c("Ljung-Box"), fitdf = 5)
LBQPlot(na.omit(residuos), lag.max = 24, StartLag = 1, k = 5, SquaredQ = FALSE)

## Normalidade dos resíduos
qqnorm(residuos)
qqline(residuos)
shapiro.test(residuos)

## Valores estimados e previsões ----

## Modelo sobre a série de retornos
# Os coeficientes são os mesmos, apenas para facilitar a elaboração das previsões
novo_modelo <- arima(ibc_var, order=c(2,0,2), seasonal = c(0,1,1),
                    include.mean = F, fixed = c(NA,NA,NA,NA,NA))
coeftest(novo_modelo)

## Desempenho do modelo na série de treinamento
plot(ibc_var,type="l",xlab="",ylab="Variação mensal")
lines(fitted(novo_modelo),type="l",lty=2,col="blue")
legend("bottomleft",
      c("Série original","Previsão"),
      lty=c(1,2),
      col=c("black","blue"))
rmserr(ibc_var, fitted(novo_modelo))

## Desempenho do modelo na série de teste
previsoes = predict(novo_modelo, n.ahead = 24, se.fit = TRUE)
conf_sup = previsoes$pred + 1.645*previsoes$se
conf_inf = previsoes$pred - 1.645*previsoes$se
ibc_vartd = ts(ibc_tudo$var, start = c(2003,2), frequency = 12)
datas_graf= time(tail(ibc_vartd,24))
ts.plot(tail(ibc_vartd,24),ylim=c(-0.13,0.13),xlab="",ylab="Variação mensal")
polygon(c(rev(datas_graf), datas_graf), c(rev(conf_sup), conf_inf),
      col = 'lightgrey', border = NA)

```

```
lines(tail(ibc_vartd,24))
lines(previsoes$pred, col="blue", lty=2)
lines(conf_sup, col="grey", lty=1)
lines(conf_inf, col="grey", lty=1)
legend("bottomleft",
      c("Série original","Previsão","Intervalo de confiança"),
      lty=c(1,2,1),
      col=c("black","blue","grey"))
rmserr(tail(ibc_vartd,24), previsoes$pred)

## Série completa com previsões
plot(ibc_vartd,xlab="",ylab="Variação mensal")
lines(previsoes$pred, col="blue",lty=2)
legend("bottomleft",
      c("Série original","Previsão"),
      lty=c(1,2),
      col=c("black","blue"))
```


APENDICE B – Código para determinação dos limites de controle do gráfico

```
## Bibliotecas adicionais
library(sarima)
library(astsa)

## Pegando os coeficientes do modelo para poder fazer a simulação
ar1 = novo_modelo$coef[["ar1"]]
ar2 = novo_modelo$coef[["ar2"]]
ma1 = novo_modelo$coef[["ma1"]]
ma2 = novo_modelo$coef[["ma2"]]
sma1 = novo_modelo$coef[["sma1"]]
desv = sd(residuos)

## Definindo os parâmetros iniciais da simulação
arl0 = 36
N = 10000
lam = 0.6

## Valores do L que vão variar para encontrar o limite de controle
L = 0
step = 0.01

fim = F
arls = c()

while (fim == F){
  soma_arl = 0
  print(L)

  ## Simulação da série temporal estacionária (ibc_des)
  for (i in c(1:N)){
    x <- sarima.sim(ar = c(ar1,ar2), d = 0, ma = c(ma1,ma2),
```

```

        sar = c(), D = 0, sma = c(sma1),
        S = 12, sd=desv, n = 1000)

z0 = mean(x)
z = c(z0)
mi = mean(x)

t = 0
fora = F

while (fora == F){
    t = t + 1
    zt = lam*x[t] + (1-lam)*z[t]
    z = append(z,zt)

    ## Cálculo dos limites de controle
    LSC <- 0 + L*desv
    LIC <- 0 - L*desv

    ## Checagem para ver se a observação está fora dos limites
    if (zt>LSC | zt<LIC) {
        soma_arl = soma_arl + t
        fora = T
    }
}

## ARL médio do L utilizado
arl_sim = soma_arl / N
arls = append(arls,arl_sim)

## Se o ARL médio for maior que o definido, a simulação termina
if (arl_sim >= arl0){
    fim = T
} else {

```

```
    L = L + step
  }

  print(arl_sim)
}

Ls = seq(from = 0, to = 1.73, by = 0.01)

## Plotagem do ARL x L, para verificar se está dentro do esperado
plot(Ls, arls, type="l", xlab="L", ylab="ARL")
```


APÊNDICE C – Código para simulação dos ARLs do processo fora de controle

```
## Definindo os parâmetros iniciais para a simulação
L = 1.73
N = 10000
lam = 0.6

## Multiplicador do desvio padrão dos resíduos, que vai variar para testar o ARL
k = 1
step = 0.05

arls = c()

while (k <= 5) {
  soma_arl = 0
  print(k)
  for (i in c(1:N)){

    ## Simulação da série temporal estacionária (ibc_des)
    x <- sarima.sim(ar = c(ar1,ar2), d = 0, ma = c(ma1,ma2),
                  sar = c(), D = 0, sma = c(sma1),
                  S = 12, mean = 0, sd=k*desv, n = 1000)
    z0 = mean(x)
    z = c(z0)

    t = 0
    fora = F
    while (fora == F){
      t = t + 1
      zt = lam*x[t] + (1-lam)*z[t]
      z = append(z,zt)

      ## Cálculo dos limites de controle para cada observação
```

```
LSC <- 0 + L*desv
```

```
LIC <- 0 - L*desv
```

```
## Checagem para ver se a observação está fora dos limites
```

```
if (zt>LSC | zt<LIC) {
```

```
  soma_arl = soma_arl + t
```

```
  fora = T
```

```
}
```

```
}
```

```
}
```

```
## ARL médio do k utilizado
```

```
arl_sim = soma_arl / N
```

```
arls = append(arls,arl_sim)
```

```
k = k + step
```

```
print(arl_sim)
```

```
}
```

```
ks = seq(from = 1, to = 5, by = 0.05)
```

```
## Plotagem do ARL x k, para verificar se está dentro do esperado
```

```
plot(ks, arls, type="l", xlab="k", ylab="ARL")
```

APÊNDICE D – Código para aplicação do gráfico de controle à série do IBC-Br

```
## Biblioteca adicional
library(ggplot2)

## Definição dos parâmetros iniciais
L = 1.73
lam = 0.6

# Pegando os valores da série Dt
z0 = mean(ibc_des)
z = c(z0)
x = ibc_des

## Contagem dos pontos fora dos limites de controle
maior = 0
menor = 0

## Cálculo dos limites de controle
LSC <- 0 + L*desv
LIC <- 0 - L*desv

for (i in c(1:length(ibc_des))){

  ## Cálculo da estatística Zt
  zt = lam*x[i] + (1-lam)*z[i]
  z = append(z,zt)

  ## Verificação se o ponto está dentro ou fora dos limites
  if (zt > LSC) {maior = maior + 1}
  if (zt < LIC) {menor = menor + 1}
}
```

```
## Plotagem da série Zt juntamente com os limites de controle
plot(z, x=c(1:181),type="l",xlab="t",ylab="Zt")
lines(rep(LIC,181),type="l",col="blue")
lines(rep(LSC,181),type="l",col="blue")
```

```
## Cálculo do ARL da série
arl_serie = length(IBC_des) / (maior+menor)
print(arl_serie)
```

```
## Série inteira, com as observações de teste
IBC_t = IBC_tudo[c(6:221),]
IBC_t_nivel = IBC_t$valor
colnames(IBC_t) = c("datas","nivel","var")
IBC_t_var = ts(IBC_t$var, start = c(2003,7), frequency = 12)
plot(IBC_t_var)
```

```
## Série diferenciada
IBC_t_des = IBC_t_var %>% diff(lag=12)
plot(IBC_t_des,type="l")
```

```
## Cálculo dos Zt para os pontos de teste
x = IBC_t_des
z_teste = z

for (i in c((length(IBC_des)+1):length(IBC_t_des))){
```

```
  ## Cálculo da estatística Zt
  zt = lam*x[i] + (1-lam)*z_teste[i]
  z_teste = append(z_teste,zt)
```

```
}
```

```
## Gráfico com os pontos de teste
plot(z_teste, x=c(1:205),type="l")
```

```

lines(rep(LSC,205),type="l",col="blue")
lines(rep(LIC,205),type="l",col="blue")

## Contagem dos pontos fora dos limites de controle
maior = 0
menor = 0
fora = c()
valores = c()

for (i in c(1:length(IBC_t_des))){

  ## Verificação se o ponto está dentro ou fora dos limites
  if (z_teste[i] > LSC) {
    maior = maior + 1
    fora = append(fora,i-1)
    valores = append(valores, z_teste[i])
  }
  if (z_teste[i] < LIC) {
    menor = menor + 1
    fora = append(fora,i-1)
    valores = append(valores, z_teste[i])
  }

}

arl_serie = length(IBC_des) / (maior+menor)
print(arl_serie)

## Plot do gráfico final de Zt
datas = seq.Date(from = as.Date("2004/7/1"), by = "month", length.out = 204)
lscs = data.frame(rep(LSC,204))
lics = data.frame(rep(LIC,204))
zdf_teste = data.frame(datas,z_teste[2:205])
zdf = data.frame(datas[1:180],z[2:181])

```

```
forad = datas[c(forad)]
pontos_fora = data.frame(forad,valores)
print(pontos_fora)
```

```
graf1 = ggplot() +
  geom_line(data = zdf_teste, aes(x = datas, y = z_teste.2.205.), color = "red") +
  geom_line(data = zdf, aes(x = datas.1.180., y = z.2.181.), color = "black") +
  geom_line(data = lscs, aes(x = datas, y = rep(LSC,204)), color = "blue") +
  geom_line(data = lics, aes(x = datas, y = rep(LIC,204)), color = "blue") +
  geom_point(data = pontos_fora[1:7,], aes(x = forad, y = valores), color = "black") +
  geom_point(data = pontos_fora[8:12,], aes(x = forad, y = valores), color = "red") +
  xlab('Datas') +
  ylab('Zt')
plot(graf1)
```

Gráfico da série estacionária com destaque para os pontos fora de controle

```
ibcdf_t_des = data.frame(datas,ibc_t_des)
ibcdf_des = data.frame(datas[1:180],ibc_t_des[1:180])
ibc_fora = ibc_t_des[c(forad)]
pontos_ibc_fora = data.frame(forad, ibc_fora)
```

```
graf2 = ggplot() +
  geom_line(data = ibcdf_t_des, aes(x = datas, y = ibc_t_des), color = "red") +
  geom_line(data = ibcdf_des, aes(x = datas.1.180., y = ibc_t_des.1.180.), color =
"black") +
  geom_point(data = pontos_ibc_fora[1:7,], aes(x = forad, y = ibc_fora), color = "black")
+
  geom_point(data = pontos_ibc_fora[8:12,], aes(x = forad, y = ibc_fora), color = "red")
+
  xlab('Datas') +
  ylab('Dt')
plot(graf2)
```

Gráfico da série de retornos com destaque para os pontos fora de controle

```
ibc_var_tudo = ibc_tudo$var[18:221]
ibcdf_var = data.frame(datas,ibc_var_tudo)
ibc_var_fora = ibc_var_tudo[c(forad)]
pontos_ibc_var_fora = data.frame(forad, ibc_var_fora)

graf3 = ggplot() +
  geom_line(data = ibcdf_var, aes(x = datas, y = ibc_var_tudo), color = "red") +
  geom_line(data = ibcdf_var[1:180,], aes(x = datas, y = ibc_var_tudo), color = "black")
+
  geom_point(data = pontos_ibc_var_fora[1:7,], aes(x = forad, y = ibc_var_fora), color
= "black") +
  geom_point(data = pontos_ibc_var_fora[8:12,], aes(x = forad, y = ibc_var_fora), color
= "red") +
  xlab('Datas') +
  ylab('Rt')
plot(graf3)
```


ANEXO A – Série original do IBC-Br

Data	IBC-Br	Data	IBC-Br	Data	IBC-Br	Data	IBC-Br
jan/03	96,15	dez/07	122,43	nov/12	144,15	out/17	135,99
fev/03	98,67	jan/08	121,86	dez/12	139,52	nov/17	135,08
mar/03	103,41	fev/08	121,91	jan/13	139,32	dez/17	135,84
abr/03	102,19	mar/08	128,99	fev/13	136,14	jan/18	132
mai/03	100,3	abr/08	129,48	mar/13	148,01	fev/18	129,85
jun/03	98,58	mai/08	128,92	abr/13	149,79	mar/18	141,63
jul/03	103,05	jun/08	130,59	mai/13	147,03	abr/18	139,05
ago/03	101,47	jul/08	136,53	jun/13	144,87	mai/18	132,27
set/03	102,87	ago/08	133,87	jul/13	152,13	jun/18	137,02
out/03	105,06	set/08	132,6	ago/13	151,81	jul/18	140,98
nov/03	101,95	out/08	132,8	set/13	147,27	ago/18	143,36
dez/03	99,74	nov/08	124,58	out/13	151,9	set/18	135,8
jan/04	98,59	dez/08	118,9	nov/13	147,79	out/18	139,77
fev/04	99,45	jan/09	115,2	dez/13	145,77	nov/18	137,59
mar/04	111,98	fev/09	115,26	jan/14	142,72	dez/18	136,32
abr/04	107,36	mar/09	127,81	fev/14	143,53	jan/19	133,41
mai/04	106,03	abr/09	123,09	mar/14	149,03	fev/19	133,76
jun/04	107	mai/09	124,41	abr/14	147,69	mar/19	138,98
jul/04	111,47	jun/09	125,61	mai/14	147,14	abr/19	139,48
ago/04	110,65	jul/09	131,44	jun/14	140,88	mai/19	139,24
set/04	109,21	ago/09	130,7	jul/14	149,85	jun/19	134,99
out/04	108,89	set/09	129,81	ago/14	148,27	jul/19	142,94
nov/04	109,59	out/09	132,48	set/14	148,12	ago/19	141,77
dez/04	107,56	nov/09	129,62	out/14	149,7	set/19	138,25
jan/05	103,52	dez/09	129,23	nov/14	144,92	out/19	142,8
fev/05	104	jan/10	125,81	dez/14	145,48	nov/19	138,82
mar/05	115,42	fev/10	127,61	jan/15	138,99	dez/19	137,5
abr/05	112,35	mar/10	143,44	fev/15	136,7	jan/20	133,72
mai/05	110,86	abr/10	136,87	mar/15	149,72	fev/20	134,77
jun/05	111,5	mai/10	136,52	abr/15	142,52	mar/20	136,84

jul/05	113,15	jun/10	136,09	mai/15	139,94	abr/20	119,25
ago/05	115,15	jul/10	141,64	jun/15	138,71	mai/20	119,91
set/05	110,95	ago/10	141,55	jul/15	143,36	jun/20	126,29
out/05	111,33	set/10	139,46	ago/15	140,94	jul/20	135,84
nov/05	111,73	out/10	139,33	set/15	138,23	ago/20	135,51
dez/05	111,25	nov/10	139,68	out/15	140,26	set/20	137,12
jan/06	108,55	dez/10	136,69	nov/15	136,03	out/20	139,74
fev/06	107,8	jan/11	132,66	dez/15	136,45	nov/20	138,1
mar/06	119,09	fev/11	136,18	jan/16	128,47	dez/20	139,41
abr/06	112,61	mar/11	144,93	fev/16	130,93	jan/21	132,03
mai/06	117,19	abr/11	139,89	mar/16	140,53	fev/21	134,8
jun/06	114,4	mai/11	143,23	abr/16	136,12	mar/21	144,06
jul/06	119,41	jun/11	141,75	mai/16	133,7	abr/21	138,91
ago/06	121,06	jul/11	145,19	jun/16	135,45	mai/21	136,85
set/06	116,21	ago/11	147,51	jul/16	136,77	jun/21	137,46
out/06	119,33	set/11	142,3	ago/16	138,23	jul/21	143,28
nov/06	118,67	out/11	142,02	set/16	134,02	ago/21	141,94
dez/06	116,3	nov/11	141,87	out/16	132,86	set/21	139,21
jan/07	114,79	dez/11	139,23	nov/16	132,51		
fev/07	113,33	jan/12	133,34	dez/16	133,55		
mar/07	125,11	fev/12	135,35	jan/17	128,57		
abr/07	120,29	mar/12	146,35	fev/17	129,56		
mai/07	123,9	abr/12	139,85	mar/17	142,05		
jun/07	122,38	mai/12	144,56	abr/17	133,75		
jul/07	127,85	jun/12	142,28	mai/17	135,95		
ago/07	129,05	jul/12	147,46	jun/17	135,05		
set/07	123,24	ago/12	149,91	jul/17	138,39		
out/07	129,17	set/12	141,6	ago/17	140,14		
nov/07	125,88	out/12	147,71	set/17	134,92		