UNIVERSIDADE FEDERAL DE MINAS GERAIS
Programa de Pós-graduação em Estatística
Especialização em Estatística Aplicada

Jonathan Enrique Pérez

EXPLORANDO A RESILIÊNCIA DO VAREJO: Uma análise preditiva com indicadores de desempenho.

Belo Horizonte

2023

Jonathan Enrique Pérez

EXPLORANDO A RESILIÊNCIA DO VAREJO: Uma análise preditiva com indicadores de desempenho.

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Especialista em Estatística Aplicada pela Universidade Federal de Minas Gerais.

Área de concentração: Estatística.

Orientadora: Profa. Dra. Ela Mercedes Medrano de Toscano.

Coorientador: Prof. Dr. Luis Alberto Toscano Medrano.

Belo Horizonte

2023

Pérez, Jonathan Enrique.

P438e Explorando a resiliência do varejo: [recurso eletrônico] uma análise preditiva com indicadores de desempenho / Jonathan Enrique Pérez – 2023.
1 recurso online (50 f. il, color.): pdf.

Orientadora: Ela Mercedes Medrano de Toscano.
Coorientadora: Luís Alberto Toscano Medrano.

Monografia (Especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 49-50.

1. Estatística. 2. Análise de séries temporais. 3. Comércio varejista – Brasil. 4. COVID-19 (Doença) – Aspectos econômicos. I. Toscano, Ela Mercedes Medrano de. II. Medrano Toscano, Luís Alberto. III. Universidade Federal de Minas Gerais, Instituto De Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas

Departamento de Estatística
P Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX:

ATA DO 324ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE JONATHAN ENRIQUE PÉREZ.

Aos quinze dias do mês de dezembro de 2023, às 10:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Jonathan Enrique Pérez**, intitulado: "Explorando a Resiliência do Varejo: Uma análise preditiva com indicadores de desempenho.", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Ela Mercedes Medrano – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 15 de dezembro de 2023.

Profa. Dra. Ela Mercedes Medrano (Orientadora)
DEST/ICEX/UFMG
Documento assinado digitalmente
gov.br LUIS ALBERTO TOSCANO MEDRANO
Data: 06/01/2024 01:06:23-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Luis Alberto Toscano Medrano
UFRRJ
Documento assinado digitalmente
gov.br MARIO ERNESTO PISCOYA DIAZ
Data: 05/01/2024 11:57:53-0300
Verifique em https://validar.iti.gov.br

Prof. Mario Ernesto Piscoya Diaz
IME/UFG

AGRADECIMENTOS

Expresso minha gratidão intensa e amorosa a minha mãe Maria Angélica Pérez, minhas irmãs e minha família na Venezuela e no Brasil, por seu apoio incondicional. Aos amigos que estiveram na trilha, oferecendo auxílio moral e positividade, meu coração transborda de amor por cada um de vocês, pois são parte essencial da minha trajetória na vida.

“Que todos nossos esforços estejam sempre focados no desafio à impossibilidade. Todas as grandes conquistas humanas vieram daquilo que parecia impossível” (autor desconhecido).

RESUMO

A indústria do varejo está passando por transformações rápidas, impulsionadas pelo crescimento do comércio online e pela crescente competição. Este estudo aborda o impacto de eventos inesperados, como a pandemia de COVID-19 em 2019, que desafiou a resiliência do setor de varejo, especialmente nas cadeias de suprimentos de produtos essenciais. Diante deste cenário, a pesquisa propõe um modelo de previsão de séries temporais utilizando indicadores de desempenho do IBGE para antecipar e responder proativamente às flutuações de mercado. Foca-se especialmente nos setores essenciais, como combustíveis, avaliando a qualidade do uso desses indicadores na previsão do desempenho varejista. O estudo visa contribuir para a literatura econômica e fornecer resultados relevantes para o planejamento estratégico e alocação de recursos no setor.

Palavras-chave: varejo; série temporal; setores essenciais.

ABSTRACT

The retail industry is undergoing rapid transformations driven by the growth of online commerce and increasing competition. This study addresses the impact of unexpected events, such as the 2019 COVID-19 pandemic, which challenged the resilience of the retail sector, particularly in the supply chains of essential products. In this context, the research proposes a time series forecasting model using performance indicators from IBGE to anticipate and proactively respond to market fluctuations. Assessing the quality of using these indicators in predicting retail performance. The study aims to contribute to economic literature and provide relevant insights for strategic planning and resource allocation in the sector.

Keywords: retail; time series; essential sectors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
IBGE	Instituto Brasileiro de Geografia e Estatística
UFMG	Universidade Federal de Minas Gerais

SUMÁRIO

1 INTRODUÇÃO

A indústria do varejo está passando por desenvolvimentos rápidos tanto na sua estrutura, com crescimento dos negócios online quanto no ambiente competitivo que as empresas enfrentam (Fildes et al, 2022). Essas mudanças estão acontecendo em um contexto em que os indicadores de desempenho econômicos desempenham um papel estratégico, permitindo que as empresas antecipem e respondam proativamente às flutuações de mercado.

No entanto, eventos inesperados, conhecidos na economia como “externalidades” que vivenciamos, por exemplo, no ano de 2019 devido à pandemia de COVID-19, apresentam desafios significativos para o setor de varejo. Isso se manifesta através do aumento da volatilidade da demanda e da ameaça à resiliência das cadeias de suprimentos, especialmente no setor de produtos essenciais (Ghafour e Aljanabi, 2023).

Desta maneira, os efeitos da pandemia de COVID-19 não apenas trouxeram desafios enormes para o setor de saúde, mas também devido às restrições relacionadas à crise sanitária, tornou o fornecimento de bens essenciais (como alimentos e vestimenta) um problema logístico significativo tanto para distribuidores quanto para varejistas (Banco Mundial, 2021). Portanto, no cenário econômico complexo que as empresas e governos atuam, a capacidade de se antecipar aos comportamentos da demanda futura é essencial para o planejamento eficaz e a alocação de recursos aprimorada.

Neste sentido, vários estudos na literatura abordaram a previsão de demanda no setor varejista. Por exemplo, Miotto e Parente (2015) ofereceram um framework para mapear o formato, a estratégia competitiva e os estágios de ciclos de modernização, permitindo aos varejistas ajustar suas variáveis de marketing. (Wang et al., 2023) desenvolveram um algoritmo de aprendizado de máquina para previsão de demanda na indústria varejista. Já Ghafour e Aljanabi (2023) focaram na resiliência das cadeias de suprimentos diante de interrupções.

Todas essas análises evidenciaram a inquietação da previsão de flutuações da demanda no setor de varejo. Como se observa há uma diversidade de 13abordagens na literatura evidenciando essa temática. Contudo poucas exploraram por meio de análises de indicadores de desempenho da produção varejista nos países como um

papel estratégico, permitindo que as empresas antecipem e respondam proativamente às flutuações de mercado.

De maneira específica, pretende-se, através da metodologia de Box e Jenkins, identificar um modelo do setor de combustível que acompanhem a variabilidade das séries e avaliar a capacidade preditiva desses modelos em relação ao uso de indicadores de desempenho varejista.

1.1 Objetivos

1.1.1 Objetivo geral

Explorar a resiliência do setor varejista frente a eventos inesperados, como a pandemia de COVID-19, por meio de um modelo de previsão de séries temporais com indicadores de desempenho da produção varejista, especificamente para o setor de combustível.

1.1.2 Objetivos específicos

Desenvolver um modelo de previsão tradicional de séries temporais utilizando indicadores de desempenho da produção varejista secundários do IBGE.

Analisar o comportamento conjuntural de dois segmentos do comércio varejista, com foco nos setores essenciais, especificamente, o setor de combustível.

1.1.3 Justificava

Um estudo dos indicadores de desempenho da produção varejista se justifica por várias razões: no âmbito da pesquisa científica, amplia a literatura econométrica e facilita as pesquisas futuras por apresentar uma visão diferente do que já é discutido sobre essa área. No âmbito de mercado, é que o desempenho do varejo restrito no Brasil em 2021, que, mesmo diante das incertezas provocadas pela pandemia, apresentou expansão de 13,9%, três vezes superior ao crescimento do Produto Interno Bruto (PIB) de 4,6% conforme o sebrae (2023). Além disso, devido à importância que se está dando ao varejo no mundo, não surpreende o interesse que o setor varejista desperta nos acadêmicos, devido ao crescente aumento de revistas

internacionais dedicadas à pesquisa sobre o varejo e edições especiais que refletem essa atenção (Mou et al, 2018).

Finalmente o estudo está organizado da seguinte maneira: na seção 2, uma revisão de estudos anteriores sobre o tema, consequentemente na seção 3, a metodologia, variáveis e modelos econométricos utilizados, na seção 4 os resultados, e por fim, na seção 5, as principais conclusões do estudo.

2. REVISÃO DA LITERATURA

Conforme Wang et al (2019), as empresas no setor de varejo tradicionalmente adotaram metodologias de previsão para orientar suas operações e produção. No passado, diversas tecnologias foram empregadas para atender às demandas dos clientes. Portanto, a previsão tradicional baseada em séries temporais, com base em análises estatísticas, tem sido uma área de pesquisa ativa ao longo das décadas.

Com isso, foram desenvolvidas abordagens estatísticas lineares, incluindo o modelo Auto-Regressivo (AR), o modelo Média Móvel (MA), o modelo Auto-Regressivo de Média Móvel (ARMA) e o modelo de Média Móvel Integrada Auto-Regressiva (ARIMA) para os modelos de cadeia de suprimentos em várias etapas (Gilbert, 2005).

Ao longo dos anos, diversas pesquisas têm contribuído significativamente para a área de previsão de demanda, demonstrando a evolução das técnicas e abordagens utilizadas nesse campo.

Em estudo realizado por Olsson e Soder (2008), a abordagem da Média Móvel Integrada Sazonal Autoregressiva (SARIMA) foi combinada com processos discretos de Markov para prever o preço do mercado de energia, considerando fatores sazonais relacionados às condições climáticas.

Além disso, Babu e Reddy (2012) realizaram uma pesquisa em que compararam a precisão de diferentes abordagens do ARIMA na previsão de temperatura em 2012, concluindo que o ARIMA baseado em tendência apresentou melhor desempenho, avaliado por meio do Erro Percentual Médio Absoluto, Erro Percentual Máximo Absoluto e Erro Médio Absoluto dos resultados.

Por outro lado, Fattah et al (2018) concentraram-se na modelagem e previsão da demanda em uma empresa de alimentos, utilizando dados históricos de demanda de alimentos e a abordagem de séries temporais de Box-Jenkins, fornecendo diretrizes confiáveis para a tomada de decisões dos gestores da empresa. Isso destacou o impacto significativo da análise de dados históricos de demanda na gestão da cadeia de suprimentos e nas estratégias de tomada de decisão.

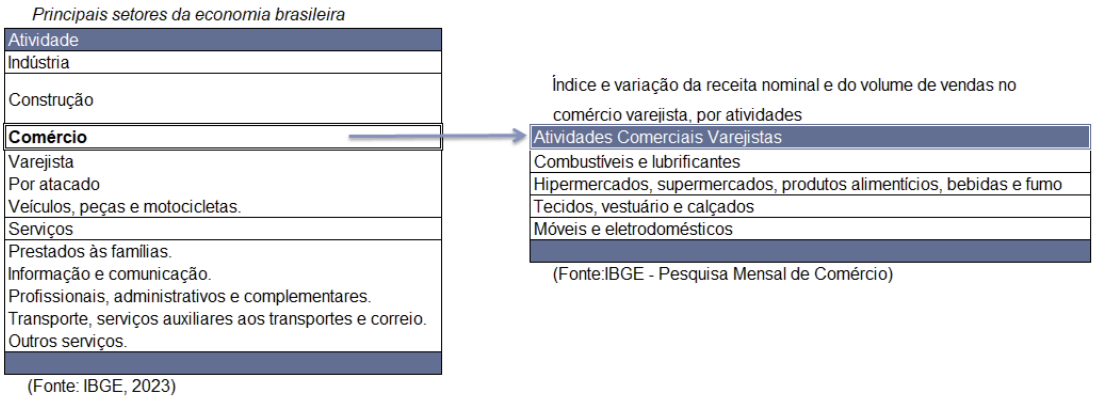
Por fim, Wang et al. (2023) desenvolveram um algoritmo de aprendizado de máquina chamado árvore de aumento de gradiente espacial-temporal (ST-BGT) para prever a demanda na indústria varejista, incorporando informações de corte transversal e séries temporais, aprimorando a capacidade de automação e a precisão

da previsão de demanda. Essas pesquisas ao longo do tempo ilustram a constante evolução e aprimoramento das técnicas de previsão de demanda, contribuindo para a eficácia na gestão de cadeias de suprimentos e estratégias de negócios.

3. METODOLOGIA

O estudo a seguir será conduzido com foco nos setores-chave da economia brasileira, concentrando-se particularmente nas atividades de comércio varejista conforme apresentado na Figura 1 a seguir:

Figura 1 - Atividades comerciais varejistas a serem trabalhadas no estudo.



Fonte: IBGE, 2023.

O período compreendido vai de Janeiro de 2000 a Maio de 2023, totalizando 281 meses por variável que ao todo representam 281 observações, pois se trabalhará exclusivamente com a variável combustível. Todos os dados foram coletados no IBGE, o site disponibiliza uma vasta gama de estatísticas sobre a realidade brasileira em vários formatos de produtos, que podem ser acessados por meio do portal IBGE na internet.

4. CONCEITOS BÁSICOS DE SÉRIES TEMPORAIS

Por definição, uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo, portanto, quando se lida com séries temporais a ordem dos dados cumprem um papel essencial. Tal instrumento é de grande relevância para as Ciências Sociais como um todo, uma vez que grandes temas abordados por essa área do saber podem ser observados utilizando uma série temporal, como taxas de mortalidade e fecundidade, fluxos de migração, dentre outros.

Formalmente, uma série temporal é definida como um processo estocástico, ou seja, uma série de variáveis indexadas no tempo, tal processo possui a seguinte notação:

$$\{Y(t,w), t \in T, w \in \Omega\}$$

em cada $t \in T, Y(t,\cdot)$ é uma variável aleatória no espaço Ω. Ou seja, para cada ponto no tempo existe um valor específico atrelado à ele, que juntos compõem o espaço Ω. Usando linguagem matemática, a série temporal observada é construída por uma sequência de números:

$$\{y\}_{t=1}^{T} = \{y_1, y_2, y_3, \dots., y_T\}$$

4.1 Decomposição Clássica

Quando se está trabalhando com séries temporais, muitas de suas propriedades podem ser captadas assumindo-se a chamada decomposição clássica, que consiste em:

$$Y_t = T_t + S_t + \varepsilon_t$$

em que Y_t é uma variável estocástica, ou aleatória, T_t é o componente de tendência da série, S_t é o componente cíclico ou sazonal e ε_t é o erro aleatório, ou seja, a parte não explicada do modelo (com média zero e variância constante).

4.1.1 Sazonalidade

O componente cíclico em questão se repete a cada intervalo de tempo fixo s, sendo assim:

$$\ldots = S_{t-2s} = S_{t-s} = S_t = S_{t+s} = S_{t+2s}$$

É a partir disso que muitas séries apresentam um comportamento que tende a se repetir em intervalos de tempo equidistantes (a cada s períodos de tempo), como é o caso de funções trigonométricas.

Existem diversas formas de eliminar a sazonalidade dos dados, mas o mais simples e direto é captar os efeitos da sazonalidade por meio de *dummies*. De modo ilustrativo, suponha-se um conjunto de dados hipotéticos com frequência trimestrais, pode-se fazer a seguinte regressão:

$$Y_t = B_0 + B_1 * D_2 + B_2 * D_3 + B_3 * D_4 + \varepsilon_t$$

isto também significa que todos os efeitos medidos por D_i serão relativos ao primeiro trimestre. Ainda poderia se usar ε_t como a série dessazonalizada, ou seja:

$$Y_t^* = B_0 + B_1 * D_2 + B_2 * D_3 + B_3 * D_4$$

$$\varepsilon_t = Y_t - Y_t^*$$

4.1.2 Tendência

Tendência é uma mudança de longo prazo no nível médio da série, ou, em termos formais:

$$Y_t = B_0 + B_1 t + \varepsilon_t$$

Sendo que o nível médio da série no tempo t é dado por: $m = B_0 + B_1 t$, com isso percebe-se que o termo de tendência é uma função determinística no tempo.

De forma semelhante ao que fora feito para remoção da sazonalidade, estima-se a seguinte equação:

$$Y_t^* = B_0 + B_1 t$$

Chamara-se de resíduos a diferença:

$$\varepsilon_t = Y_t - Y_t^*$$

Assim sendo, tem-se a série original, menos a parte que capta os efeitos da tendência.

4.2 Estacionariedade

4.2.1 Processos Estocásticos Estacionários

O conceito de estacionariedade se divide em fracamente estacionário (ou apenas estacionário) e estritamente estacionário. O primeiro perpassa por um processo estocástico se, e somente se este for:

1. $EY(t) = \mu(t) = \mu, \forall t \in T$
2. $EY^2(t) < \infty, \forall t \in T$
3. $\gamma(t_1, t_2) = Cov\big(Y(t_1), Y(t_2)\big)$ é $uma\ função\ apenas\ de\ |t_1 - t_2|$

Por outro lado, um processo estocástico é estritamente estacionário se todas as suas distribuições permanecem as mesmas sob translações no tempo.

4.2.2 Processos Estocásticos Não-Estacionários

Séries não estacionárias têm uma tendência que podem ter uma natureza determinística ou estocástica. A série não estacionária determinística acrescida de um componente aleatório (erro ou ε extraído de uma distribuição normal, flutua em torno de uma tendência temporal).

$$y_t = c + \delta t + \varepsilon_t$$

Por outro lado, a série com tendência estocástica move-se em torno de médias flutuantes.

$$y_t = c + y_{t-1} + \varepsilon_t$$

Em outras palavras, as séries determinísticas, são aquelas cuja tendência de crescimento é sempre a mesma, ou seja, a taxa média de crescimento (ou decrescimento) dos valores tende a ser um valor fixo. Por outro lado, a série com tendência estocástica é aquele cuja média da taxa média de crescimento (ou decrescimento) varia, no exemplo anterior $y_t = c + y_{t-1} + \varepsilon_t$ vemos que o valor de y_t depende do valor de y_{t-1} não sendo, portanto, um valor fixo.

Com isso, pode-se dizer que a análise de séries temporais tem objetivos variados, como compreender o mecanismo gerador da série, realizar previsões futuras, descrever seu comportamento ao longo do tempo e identificar possíveis padrões periódicos nos dados. Para alcançar esses objetivos, recorre-se a modelos estocásticos, ou seja, processos controlados por leis probabilísticas. Conforme

Morettin e Toloi (1987), esses modelos devem ser concebidos de forma simples e parcimoniosa, com o menor número possível de parâmetros.

A realização de estimativas em séries temporais demanda a suposição de que a série seja estacionária. Em termos gerais, séries econômicas são compostas por três elementos: a tendência, o componente estacionário e o ruído (Bueno, 2008).

Em termos simples, um processo estocástico é considerado estacionário quando sua média e variância permanecem constantes ao longo do tempo, e o valor da covariância entre dois pontos no tempo depende apenas da diferença ou do intervalo entre esses dois pontos, e não do tempo real em que a covariância é calculada.

Portanto, uma série é considerada estacionária fraca quando sua média, variância e covariância permanecem constantes.

$$E(Y_t) = \mu$$

$$Var(Y_t) = E(Y_t - \mu)^2$$

$$Cov\left(Y_t - Y_{t-j}\right) = E\left(Y_t - Y_{t-j}\right) - E(Y_t)E\left(Y_{t-J}\right) = Y_K$$

Séries que não exibem alguma dessas propriedades são classificadas como não estacionárias e requerem uma transformação. Geralmente, a transformação mais comum envolve a aplicação de diferenças sucessivas até que a série se torne estacionária. A determinação da estacionariedade ou não da série é realizada por meio de testes de raiz unitária[1].

Existem diversos modelos disponíveis para análise de séries temporais. Conforme Morettin e Toloi (1987), podemos categorizá-los em dois grupos principais: modelos não-paramétricos, que envolvem um número infinito de parâmetros, e modelos paramétricos, que possuem um número finito de parâmetros. Entre os modelos paramétricos, os mais comuns incluem os modelos de regressão, os modelos autorregressivos e de médias móveis (ARMA), os modelos autorregressivos integrados de médias móveis (ARIMA), modelos de memória longa (ARFIMA), modelos estruturais e modelos não-lineares.

De acordo com Morettin e Toloi (1987), a classe dos Modelos ARIMA tem a

[1] Uma raiz unitária é uma característica que indica a não estacionariedade de uma série temporal. Conforme Gujarati e Porter (2011), em um modelo de passeio aleatório, onde Yt = ρYt-1 + ut para -1 ≤ ρ ≤ 1, a presença de ρ = 1 leva à situação de não estacionariedade. Para ilustrar, ao usar o operador de defasagem L; (1-L)Yt = ut, se conseguirmos L=1, isso é chamado de raiz unitária.

capacidade de descrever o comportamento de séries estacionárias e:

a) Descrever o comportamento de séries econômicas onde os erros são autocorrelacionados e influenciam a evolução do processo.

b) Descrever séries não estacionárias do tipo homogêneas, ou seja, aquelas que não exibem comportamento explosivo.

Normalmente, séries podem ser transformadas em estacionárias com um número finito de diferenças, geralmente uma ou duas.

A primeira diferença de Y_t é definida por:

$$\Delta Y_t = Y_t - Y_{t-1,}$$

A segunda diferença é dada por:

$$\Delta^2 Y_t = \Delta[Y_t - Y_{t-1}],$$

Ou seja,

$$\Delta^2 Y_t = \Delta[Y_t - 2Y_{t-1} + Y_{t-2}],$$

De modo geral, a enésima diferença de Y_t (n≥1) é definida como:

$$\Delta^n Y_t = \Delta[\Delta^{n-1} Y_{t-1}]$$

É fundamental considerar um conceito essencial em processos estocásticos, análises e previsões de séries temporais: o ruído branco, também conhecido como um processo puramente aleatório é caracterizado por termos de erro que possuem média zero, variância constante e são não correlacionados. Além disso, se esses termos de erro seguem uma distribuição normal, o processo é denominado ruído branco gaussiano.

$$E(\varepsilon_t) = 0,$$
$$E(\varepsilon^2{}_t) = \sigma^2,$$
$$E(\varepsilon_t \varepsilon_{t-j}) = 0$$

Nesse contexto, podemos afirmar que um processo do tipo ruído branco é um exemplo de processo estocástico estacionário.

5. MODELOS PARA SÉRIES TEMPORAIS

5.1 Processos Auto Regressivos Estacionários

Um processo estocástico auto regressivo pode ser modelado por um modelo auto regressivo de ordem p. De acordo com esse modelo, uma série temporal y_t é descrita apenas por seus valores passados e por um erro normalmente distribuído, centrado em zero (0) com variância constante ε.

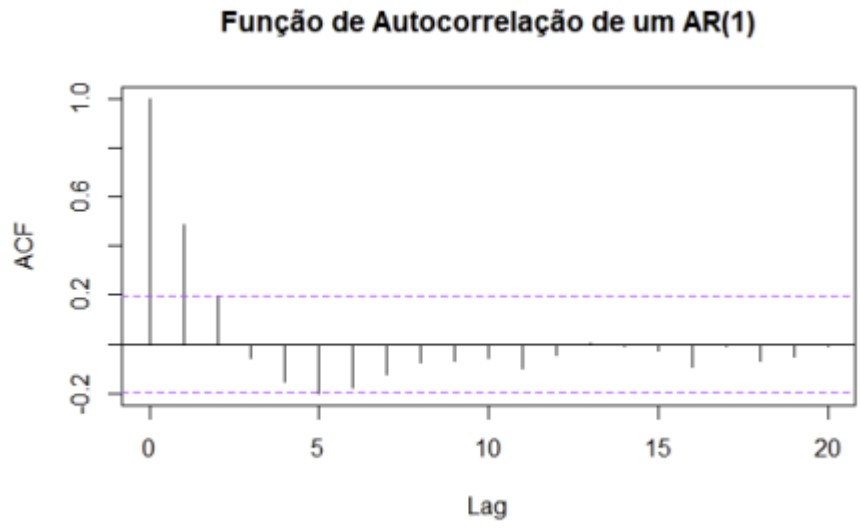
A versão mais simples de um modelo AR é aquela em que y_t depende somente de y_{t-1} e de ε_t. Diz-se, nesse caso, que o modelo é auto regressivo de ordem um (1), o que se indica abreviadamente por AR (1).

$$y_t = \phi y_{t-1} + \varepsilon_t$$

5.1.1 Função de Auto correlação

A função de auto correlação Figura 2 é uma medida de correlação de duas observações distintas, separadas por s unidades de tempo (y_t, y_{t-s}). Ou seja, ela apresenta um padrão gráfico em que vemos em cada coluna a correlação entre a variável y_t e uma de suas demais observações defasadas no tempo.

Figura 2 – Função de autocorrelação de um modelo AR(1)



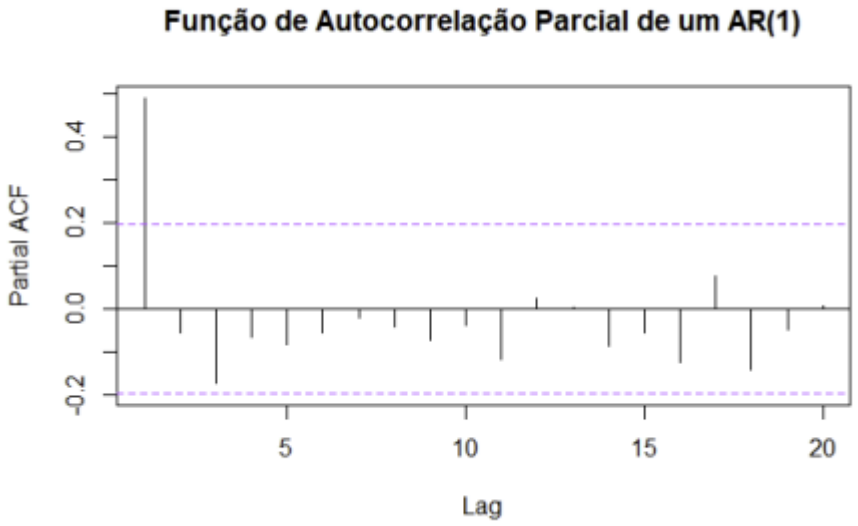
Dito, isso se observa que a primeira coluna tem altura um (1), ou seja, a correlação de y_t e y_t é, obviamente, um (1). Já a correlação entre y_t e y_{t-1} é de

aproximadamente 0.5, enquanto a correlação de y_t e y_{t-2} é aproximadamente 0.2, e assim sucessivamente.

5.1.2 Função de Auto correlação Parcial

Por outro lado, a função de autocorrelação parcial determina a ordem p do processo autoregressivo AR(p)) Figura 3.

Figura 3 – Função de autocorrelação parcial de um AR(1)



Como já enunciado, se percebe que há significância (ou seja, a coluna em questão está acima do tracejado azul) na primeira lag, demonstrando que se trata de um processo AR(1), como já fora mencionado.

A FAC e FACP possibilitam identificar o tipo de processo que descreve uma série temporal.

A autocorrelação entre Y_t e Y_{t-1} – FAC – pode ser definida como:

$$P_k = \frac{E[(Y_t - \mu)(Y_{t-1} - \mu)]}{\sqrt{[E(Y_t - \mu)^2 E\left(Y_{t-j} - \mu\right)^2]}}$$

Considerando a condição de estacionariedade, tem-se $E(Y_t - \mu)^2 = E\left(Y_{t-j} - \mu\right)^2$. Nesse sentido, $\sqrt{[E(Y_t - \mu)^2 E\left(Y_{t-j} - \mu\right)^2]} = \sigma^2 = \gamma_o$, e a autocorrelação de orde j pode ser expressa por:

$$P_k = \frac{\gamma_k}{\gamma_o}$$

Assim, pode ser apresentar todas as autocorrelações P_k para k=1,2,... em um gráfico chamado correlogramas. De acordo com Bueno (2018), “A função de autocorrelação é o gráfico de autocorrelação contra defasagem e permitirá identificar a ordem q de um processo MA”. A FAC possui a seguintes propriedades.

$$P_0 = 1$$

$$|P_j| \leq 1 \text{ para todo j}$$

$$P_{-j} = P_j \text{ para todo j}$$

Por sua vez, a Função de Autocorrelação Parcial (FACP) mantém apenas correlação pura entre as observações, eliminando as correlações implícitas (Bueno, 2008). Para Gujarati e Porter (2011), “a autocorrelação parcial” é a correlação entre Y_t e Y_{t-1} depois de remover o efeito dos Y intermediários. A FACP pode ser definida como segue:

$$Y_t = \emptyset_{j,1} Y_{t-1} + \emptyset_{j,2} Y_{t-2} + \cdots + \emptyset_{j,j} Y_{t-j} + \varepsilon_t \quad j = 1,2,3,..$$

$$\emptyset_{jj} = Corr(Y_t, Y_{t-j}/Y_{t+1}, \ldots, Y_{t+j-1})$$

5.2 Processo Autoregressivo Não Estacionário

O processo autoregressivo não estacionário é aquele cuja média ou variância não serão constante, ou seja, dependentes do tempo.

Dessa forma, para remover a tendência determinística, deve-se estimar y_t contra o tempo e armazenar os resíduos. Os resíduos armazenados constituem uma nova série que deverá ser modelada de forma separada. Por outro lado, para remover uma tendência estocástica, basta aplicar diferenciação na série.

Diante disso, chega-se ao conceito de processo estocástico integrado, que é aquele que se torna estacionário por meio de diferenciação. Dizemos que se um processo é integrado de ordem um (1) (ou I (1)) quando esse se torna estacionário na d-ésima diferença, dizemos que é um processo integrado de ordem d, ou I (d), assim como um processo não diferenciado estacionário é de ordem zero (0).

Após analisar e verificar a estacionariedade da série, o próximo passo consiste em escolher o modelo que melhor se ajusta aos dados. Em termos gerais, os modelos estudados neste trabalho são considerados casos específicos de um modelo de filtro

linear. Conforme explicado por Morettin e Toloi (1987), esses modelos pressupõem que a série temporal seja gerada por meio de um filtro, no qual a entrada é um ruído branco.

5.3 Processo Autoregressivo AR(1)

O processo autoregressivo de ordem 1- AR(1) é identificado com a seguinte estrutura:

$$Y_t = c + \emptyset Y_{t-1} + \varepsilon_t,$$

ε_t , é um ruído branco.

O valor previsto de Y no período t é uma proporção (∅) – limitado entre -1 e 1 – do seu valor no período anterior (t-1), acrescido de um choque aleatório, ou perturbação no período – t(ε). Os valores de Y são expressos como desvios com base em um valor médio. Caso o processo também dependa da sua variável defasada em dois períodos Y_{t-2}, o processo é conhecido como Autoregressivo de Ordem 2, AR(2).

Um modelo autoregressivo de ordem p – AR(p) – é definido da seguinte forma:

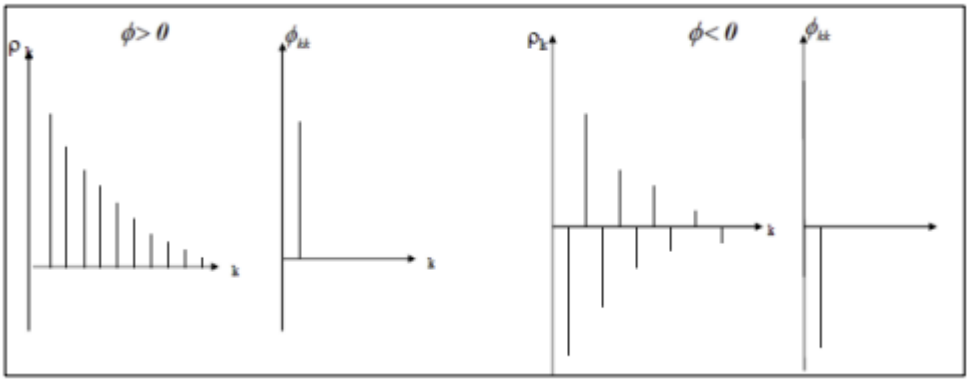
$$Y_t = c + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_p Y_{t-p} + \varepsilon_t$$

Usando um operador de defasagem B^2:

$$(1 - \emptyset_1 B - \emptyset_2 B^2 - \emptyset_3 B^3 - \cdots \emptyset_P B^P) Y_t = c + \varepsilon_t$$

O processo AR e sua ordem serão identificado pela análise da sua Função de Autocorrelação (FAC) e Função de Autocorrelação Parcial (FACP). No processo AR(1) a FAC apresenta decaimento exponencial para zero, enquanto que na FACP a primeira correlação é significativa e as demais apresentam rápido decaimento. Esse processo pode ocorrer de forma alternada, conforme apontado na Figura 4.

Figura 4 - Exemplo de correlogramas FAC e FACP modelo AR(1)



Para que um modelo AR seja estacionário, existem algumas restrições para os parâmetros.

$$\text{Para } P = 1, -1 < \emptyset < 1$$

$$\text{Para } P = 2, \emptyset_1 + \emptyset_2 < 1, \emptyset_2 - \emptyset_1 < 1 \; e \; -1 < \emptyset_2 < 1$$

Para p>2 as condições são mais complicadas, sendo necessária uma análise mais aprofundada (Bueno, 2008).

5.4 Processo Média Móvel MA(1)

O processo média móvel de ordem 1 – MA(1) é identificado com a seguinte estrutura:

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

Onde: $\varepsilon_t \sim i.i.d^2$, é um ruído branco.

Nesse processo, Y_t depende do erro presente (ε_t) e uma proporção do erro indiretamente anterior ($\theta\varepsilon_{t-1}$). Caso o processo também dependesse do erro em dois períodos anteriores – ($\theta\varepsilon_{t-2}$), seria um MA(2) assim por diante (Bueno, 2008).

A generalização do processo de médias móveis de ordem (q) - MA(q) pode ser descrito da seguinte maneira:

$$Y_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q},$$

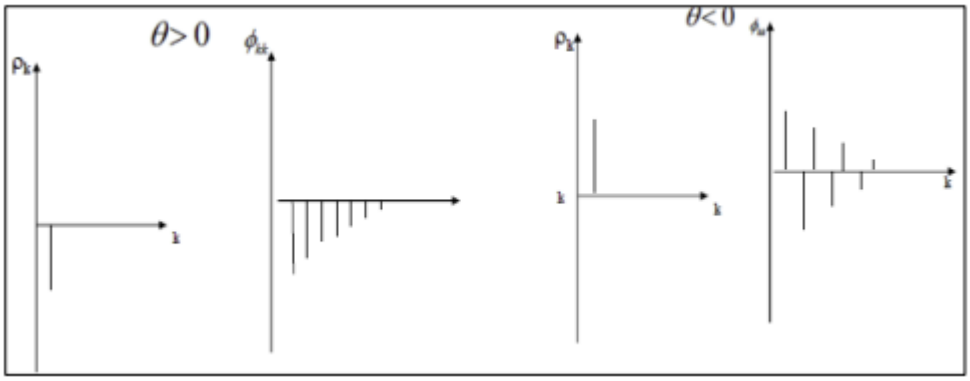
Utilizando um operador de defasagens B[.], pode se escrever:

$$Y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t$$

A identificação de um processo MA também é realizada pela análise da FAC e FACP. No processo MA(1) a FAC possui a primeira defasagem significativa, enquanto

a FACP apresenta um decaimento exponencial nas autocorrelaçoes parciais. Esse decaimento exponencial também pode ocorrer de forma alternada, conforme apontado na Figura 5.

Figura 5 - Exemplo de correlogramas FAC e FACP modelo MA(1)



Assim como o modelo AR, o processo MA também exige algumas restrições nos parâmetros. Essa condição de invertibilidade significa a capacidade de escrever uma MA(q) como um AR(∞).

$$\text{Para } q = 1, -1 < \theta < 1$$

$$\text{Para } q = 2, \theta_1 + \theta_2 < 1, \theta_2 - \theta_1 < 1 \; e \; -1 < \theta_2 < 1$$

Para q>2 as condições são mais complicadas, sendo necessária uma análise mais aprofundada (Bueno, 2008).

Ainda de acordo com Bueno (2008), a invertibilidade é necessária para três propósitos.

a) Sem a invertibilidade a série não poderia ser estimada recursivamente, usando observações passadas;
b) Para haver unicidade de resultados;
c) Para gerar funções de autocorrelação parciais.

Vale destacar que um processo de médias móveis é sempre estacionário, ao passo que processos autoregressivos são sempre invertíveis.

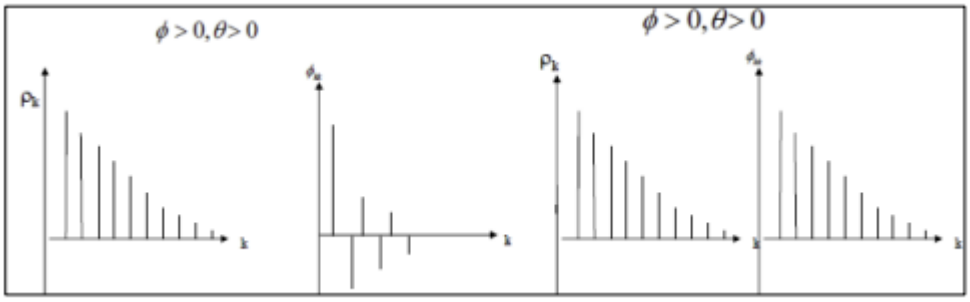
5.5 Modelos Autoregressivos e de Médias Móveis – ARMA

O processo ARMA é aquele que combina o processo autoregressivo e de médias móveis. O processo ARMA(1,1) pode ser definido como:

$$Y_t = c + \emptyset Y_{t-1} + \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

O processo é estacionário se -1<∅<1 é invertível se -1<θ<1. O processo é caracterizado por decaimento exponencial na FAC e na FACP, com autocorrelações significativas nos dois, conforme a Figura 6.

Figura 6 - Exemplo de correlogramas FAC e FACP modelo ARMA(1,1)



A generalização do processo de autoregressivos de médias móveis é o ARMA(p,q). Ele pode ser descrito da seguinte forma:

$$Y_t = c + \emptyset Y_{t-1} + \cdots \emptyset_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_p \varepsilon_{t-p}$$

Utilizando o operador B[.], pode-se reescrever um processo ARMA(p,q).

$$(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p Y_t = c + (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\varepsilon_t$$

Para que um processo ARMA(p,q) seja invertível e estacionário a parte AR precisa ser estacionária e a parte MA invertível.

Quando um processo não estacionário que exige diferenciação até que se torne estacionário, a notação dos processo ARMA será (p,d,q), sendo d a ordem de integração, ou diferenciação aplicadas na série original. Esses processos são conhecidos como “processos lineares não-estacionário homogêneos”. Após a aplicação das diferenças, o processo torna-se autoregressivo integradode médias móveis – ARIMA(pd,q).

De acordo com Epitânio (2015), além da tendência, as séries podem apresentar componentes sazonais, que repetem o comportamento em determinados períodos, por exemplo, o aumento do desemprego no primeiro trimestre do ano em função das demissões de trabalhadores temporários no final do ano anterior. Se plotarmos essa série, poderemos observar “picos” nesses períodos (s=4). O modelo pode então ser generalizado para um modelo SARIMA (p,d,q) (P,D,Q)s. As três últimas ordens referem-se a parte sazonal da série, mais especificamente, um processo autoregressivo sazonal (P), as diferenças sazonais (D) e um processo de média móvel sazonal (Q).

Em processos sazonais, a diferenciação também deve ser realizada na parte sazonal, encontrando valores para d e D.

Conforme o autor, as ordens dos processos também serão definidas na FAC e FACP. Desse modo, será necessário verificar as autocorrelações significativas em ambos, estimando os valores de p e q por meio da análise de comportamento das autocorrelações nos “lag” (defasagens) 1,2,3...e estimando os valores de P e Q por meio do comportamento da FAC nos “lags” sazonais.

5.6 Modelo SARIMA

Os modelos de classe SARIMA são uma extensão dos modelos ARIMA. Eles incorporam um componente sazonal que não é tratado pelos modelos ARIMA tradicionais. Isso é feito adicionando alguns parâmetros no modelo. A nomenclatura é SARIMA $(p,d,q)(P,D,Q)$, onde P,D,Q são parâmetros sazonais de um processo autoregressivo (P), diferenças sazonais (D), e média móveis sazonais (Q). Considerando uma série hipotética com sazonalidade de 12 períodos escrita como:

$$Y_t = \mu_t + N_t,$$

Onde μ_t é uma função determinística periódica que satisfaz $\mu_t - \mu_{t-12} = 0$ ou

$$(1 - B^{12})\mu_t = 0$$

Sendo B^{12} a série Y_t em 12 períodos anteriores, isto é, Y_{t-12}, e N_t um processo estacionário que pode ser modelado por um ARMA(p,q). Assim, N_t satisfaz

$$\alpha(B)N_t = \beta(B)\mu_t ,$$

Onde μ_t é ruído branco gaussiano. Aplicando a diferença sazonal $(1-B)^{12}$ à expressão $Y_t = \mu_t + N_t,$ chega-se a:

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

A qual se transforma em:

$$\alpha(B)W_t = \beta(B)(1 - B^{12})\mu_t$$

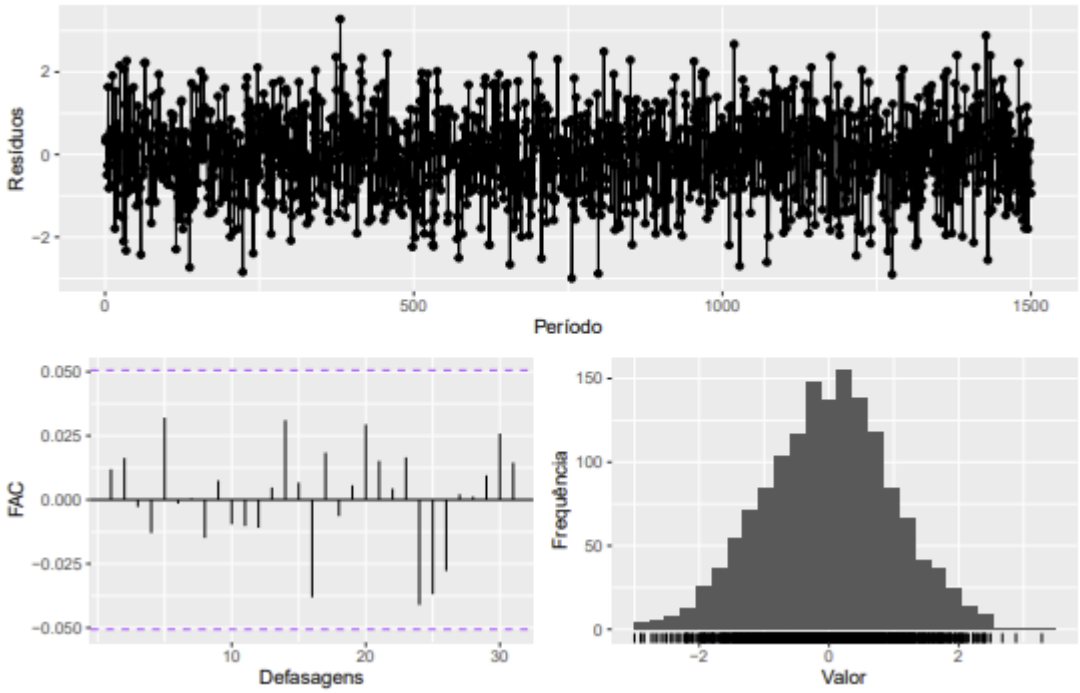
Onde $W_t = (1 - B^{12})Y_t$.

Para realizar a estimação de um modelo SARIMA, é importante chegar se a série é estacionária. Caso necessário, aplicar a diferenciação e a diferenciação sazonal. Com isso, obtêm-se os valores de d e D. Em seguida, inspeciona-se os gráficos da FAC e FACP amostrais da série para identificar os valores de P e Q

iniciais. Assim, formula-se um modelo SARIMA inicial para descrever a série.

Após estimação do modelo, é importante verificar algumas suposições dos resíduos. Em geral, assume-se que os resíduos μ_t seguem uma distribuição normal, isto é $\mu_t \sim N(0, \sigma_\mu^2)$. Visualmente, é possível verificar isso por meio de gráficos de resíduos como ilustrado na figura 5 a seguir, onde o primeiro gráfico mostra a série, o segundo plota a FAC e o terceiro o histograma. Neste caso, por tratar-se de uma série de ruído branco que será utilizada como exemplo, não há autocorrelação entre as defasagens dos resíduos e o histograma visualmente segue o formato de uma distribuição normal. Uma outra forma é aplicar os testes de raiz unitária de Kwiatkowski et al. (1992) e Dickey e Fuller (1979) para checar estacionariedade, assim como os testes de normalidade e independência utilizando os resíduos do modelo ajustado.

Figura 7 – Gráfico dos resíduos, FAC e histograma para um ruído branco gaussiano simulado pelo autor.



5.7 A metodologia Box-Jenkins

A metodologia foi introduzida em 1976 por George Box e Gwilym Jenkins, no livro *Time séries Analysis: Forecast and control.* Para rocha (2022), a principal

concepção dos modelos de Box-Jenkins é que um processo estocástico pode ser explicado pelos valores passados da série e pelo termo de erro.

Por meio da metodologia Box-Jenkins é possível identificar se o modelo segue um processo autoregressivo (AR) e, se sim, qual ordem de p; se o modelo segue um processo de médias móveis (MA), e qual o valor de q; se segue um processo Autoregressivo de Médias Móveis (ARMA) ou mesmo um processo Autoregressivo Integrado de Médias Móveis (ARIMA) e qual ordem de p,d,q.

A metodologia consiste em (4) etapas:

1. Identificação: através do auxílio dos correlogramas de função de autocorrelação (FAC)e função de autocorrelação parcial (FACP), pode se identificar os valores apropriados de p,d,q.
   a) Modelo AR(p);
   b) Modelo MA(q);
   c) Modelo ARMA (p,q).

2. Estimação: após identificar as ordens do modelo, o passo seguinte é a estimação dos parâmetros. Ela pode ser realizada pelo Método dos Momentos, Mínimos Quadrados Ordinários, Máxima Verossimilhança Condicional e Máxima Verossimilhança Não condicional.

3. Verificação: após escolher o modelo específico e estimado seus parâmetros, verifica-se se o modelo se ajusta bem à variabilidade da série.

Equação do modelo ARMA:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p Y_t = c + (1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q)\varepsilon_t$$

Equação do modelo SARIMA

$$\Phi_p(B^s)\phi_p(B)W_t = \theta_q(B)\Theta_q(B^s)a_t$$

Onde:

$W_t = (\Delta^d \Delta^D X_t), são\ as\ raízes\ dos\ polinômios;$

$\phi_p(B) = 0,$ é $o\ polinômio\ autoregressivo;$

$\Theta_q(B) = 0,$ é $o\ polinômio\ de\ médias\ móveis\ de\ ordem\ q\ com\ raízes\ com\ raízes\ fora\ do\ circulo\ unitário\ e\ sem\ raízes\ comuns;$

$\Phi_p(B^s) = 0,$ é $o\ polinômio\ autoregressivo\ sazonal\ de\ ordem\ p;$

$\Theta_Q(B^s) = 0,$ é $o\ polinômio\ de\ médias\ móveis\ sazonal\ de\ ordem\ Q\ e$

$\{a_t\}$ é $um\ processo\ Ruído\ Branco\ com\ média\ zero\ E(a_t) = 0\ e$

$$Variância\ constante\ VAR(a_t) = \sigma^2.$$

5.8 Verificação de modelos ARMA (p,q)

A verificação do modelo consiste em avaliar os seguintes passos:

a) Os parâmetros $\emptyset_s$ estão dentro da região de estacionariedade?

b) Os parâmetros $\emptyset_s$ estão dentro da condição de invertibilidade?

c) Os parâmetros são significativamente diferentes de zero?

d) O resíduo é ruído branco?

Para determinar o comportamento estacionário da série, pode-se realizar o teste de raiz unitária de Dickey-Fuller Aumentado (ADF), no qual são se obtém a estatística de Dickey – Fuller. Esse valor é comparado ao valor tabelado de Dickey – Fuller. No teste, considera-se:

H_0: $\delta = 0$ (Há raiz unitária, portanto a série é não estacionária)

H_1: $\delta < 0$ (A série é estacionária)

Por sua vez, o componente de médias móveis (MA) de uma série é sempre estacionário por definição, como já mencionado.

É preciso testar também se os parâmetros do modelo são significativamente diferentes de zero:

$$H_0: \emptyset_s = 0\ vs\ H_1: \emptyset_s \neq 0$$

$$H_0: \theta_s = 0\ vs\ H_1: \theta_s \neq 0$$

Por fim, é necessário verificar se os resíduos do modelo são do tipo ruído branco, como mencionado. Para tanto:

1. Os resíduos devem estar localizados ao redor da reta centrada e não exibe a presença de nenhuma configuração especial;

2. Os resíduos padronizados estão dentro da faixa entre -2 e 2, não sendo observados elementos discrepantes;

3. No gráfico de FAC e FACP dos resíduos, é possível validar a suposição de que os erros são não autocorrelacionados;

4. É necessário verificar ainda se os resíduos seguem uma distribuição normal, o que pode ser feito por meio do teste de Anderson Darling em que:

$$H_0: os\ erros\ seguem\ uma\ distribuição\ normal$$

$$H_1: os\ erros\ NÃO\ seguem\ uma\ distribuição\ normal$$

5.9 Escolha dos modelos

Existem muitos critérios de informação para a seleção do melhor modelo estimado, entre eles, os mais utilizados são:

1. Critério de AIC: Introduzido em 1971 por Akaike, esse é critério definido por: AIC(k) = $n\ln(\sigma_\alpha{}^2) + 2k$, sendo k o número de parâmetros estimados e $\sigma_\alpha{}^2$ a variância dos resíduos (Toscano, 2022). Por esse critério, o melhor modelo é aquele que apresenta o menor valor de AIC.
2. Critério de BIC: o Critério Bayesiano de Schwarz é dado por: BIC(p,q) = $\ln(\sigma^2)$ + (p+q) [$\frac{\ln(n)}{n}$]. Por esse critério, o melhor modelo também é aquele que minimiza o valor de BIC.

5.10 Previsão

A previsão consiste em realizar a projeção de até h passos à frente para os modelos selecionados nas etapas anteriores (Arrais, 2019).

Para escolha do melhor modelo de previsão, entre os modelos selecionados devem-se considerar algumas estatísticas baseadas nos erros de previsão. Isso porque, de acordo com Campos e Clemente e Cordeiro (2006), modelos que apresentem as evidências estatísticas que tornem consistentes, como AIC e BIC, podem não gerar resultados de previsão satisfatórios.

Com o objetivo de testar a acurácia entre os modelos propostos, são considerados alguns índices de desempenho no período amostral e no período de validação. São consideradas as estatísticas básicas dos erros:

1. Erro Percentual Absoluto Médio ou Desvio Absoluto Médio:

$$\text{MAPE} = (\frac{1}{h}\sum_{j=1}^{T+h}\frac{e_j}{y_j})100\%$$

2. Erro Quadrado Médio ou Desvio Quadrado Médio:

$$\text{MSE} = \frac{1}{h}\sum_{j=T+1}^{T+h} e^2\, j$$

3. Erro Absoluto Médio ou Desvio Absoluto Médio:

$$\text{MAD} = \frac{1}{h}\sum_{j=T+1}^{T+h}|e^2\, j|$$

4. Raiz do Erro Quadrado Médio:

$$\text{RMSE } \sqrt{\frac{1}{h}\sum_{j=T+1}^{T+h} e^2\, j}$$

A escolha do melhor modelo deve levar em consideração aquele que apresente o menor valor em algumas das estatísticas dos erros de previsão citada acima.

5.11 Análise dos resíduos

Após estimação dos modelos, é necessário verificar os resíduos. Devem-se confirmar as suposições de que os erros são do tipo ruído branco, com média zero $E[\varepsilon_t]$=0 e variância constante $Var(\varepsilon_t) = \sigma^2{}_\varepsilon$, conforme mencionado anteriormente.

Para confirmar as suposições deve-se observar na análise gráfica se os resíduos estão dispersos ao redor de uma reta centrada em zero e entre a faixa [-2,2].

Consequentemente depois dessa primeira verificação, serão analisados os correlogramas FAC e FACP dos resíduos para identificar se as correlações são estatisticamente diferentes de zero, com isso as observações deverão estar abaixo da faixa de significância. Cabe destacar que confirmadas essas hipóteses, haverá indícios de que os erros sejam do tipo ruído branco. Conforme Bueno (2008) se houver correlações significativas dos resíduos deverá se descartar o modelo estimado, pois a previsão poderá não ser assertiva.

Outra maneira que pode ser confirmada a hipóteses de autocorrelação dos resíduos é através do teste: o de Ljung-Box e descrito a seguir:

5.12 Teste de Ljung-box

Para verificar a hipóteses de não autocorrelação, ou seja, se as k primeiras autocorrelações são nulas, como segue:

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0$$

Pode se utilizar a estatística Q de Box e Pierce: $Q = \frac{T(T+2)\sum_{t-1}^{T} r_t^2}{(T-1)} \sim \chi^2(k-m)$, sendo m o número de parâmetros livres do modelo. Se Q for grande quando comparado a um percentil apropriado, rejeita-se H_0.

Em 1978, Ljung-Box propuseram uma variação do teste, a saber:

$Q(k) = n(n+2)\sum_{j=1}^{K} \frac{r^2{}_j}{(n-j)}$ que terá uma distribuição χ^2 com K-p-q graus de liberdade. A hipótese de ruído branco seria rejeitada para valores grandes de Q(K).

Ainda de acordo com Morettin e Toloi (1987), basta utilizar as 10 ou 15 primeiras defasagens.

5.13 Teste de Normalidade de Anderson-Darling

Após verificar a autocorrelação dos resíduos, deve-se verificar se eles possuem distribuição normal. Essa análise pode ser feita por meio do gráfico de probabilidade normal, que consiste na dispersão dos resíduos contra os valores ajustados (y) da série. Se o gráfico de probabilidade normal se mostrar como uma linha reta, aproximadamente, pode-se inferir que os resíduos seguem uma distribuição normal Guajati e Porter (2011).

Para além do gráfico de probabilidade normal, há o teste de Anderson-Darling. Nesse teste, a hipótese nula é a de que os erros seguem uma distribuição normal. Se o valor p da estatística AD calculada for alto, não se rejeita a hipótese nula.

Destacando que se os erros forem do tipo ruído branco, com média zero $\boldsymbol{E}[\boldsymbol{\varepsilon_t}]$=0 e variância constante $\boldsymbol{Var}(\boldsymbol{\varepsilon_t}) = \boldsymbol{\sigma^2}_{\boldsymbol{\varepsilon}}$, e seguirem uma distribuição normal, esse processo é chamado de ruído branco gaussiano conforme Toscano (2021).

6. ANÁLISE DOS RESULTADOS

6.1. Modelo para Série Combustível

Iniciara-se com a segmentação das observações em dois períodos distintos:

- Período Amostral (Azul): Este período abrange de janeiro de 2000 a maio de 2022, totalizando 269 observações;
- Período de Validação (Vermelho): Compreende o intervalo de junho de 2022 a maio de 2023, com um total de 12 observações.

Essa divisão na Figura 8 proporciona uma análise mais minuciosa, permitindo destacar o comportamento da série temporal em dois momentos distintos. É relevante ressaltar que o período amostral servirá como alicerce para a realização de previsões no período de validação, consolidando, assim, a capacidade de assertividade do modelo escolhido conforme o Gráfico 1.

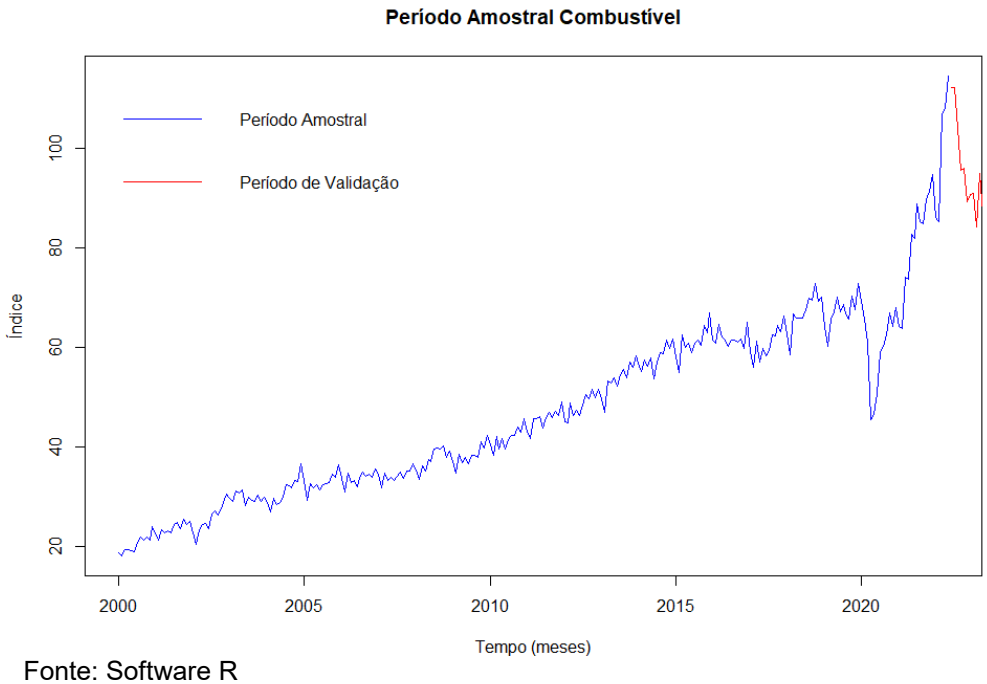
Para se identificar o modelo S(2,1,1)(0,1,1)[12] foi realizada a visualização gráfica das correlações. A identificação será realizada, essencialmente, na comparação da FAC e FACP da série original do período amostral.

Opta-se por não realizar algum tipo de transformação na série e continuar com ela em seu estado “original”. Isso se deve porque realizando algum tipo de transformação perdia-se qualidade no processo de previsão.

Na identificação inicial não se prestará atenção particular a pequenos pormenores, mas sim ao comportamento em geral, uma vez que se pode refinar o modelo numa fase posterior de avaliação do diagnóstico.

. Ao realizar uma análise, é possível obter uma compreensão abrangente do comportamento da série temporal, incluindo a identificação de tendências, sazonalidades, valores atípicos, variações na variância, entre outros aspectos. Essa representação gráfica proporciona insigths valiosos sobre a dinâmica subjacente da sucessão dos dados.

Figura 8 - Gráfico: Período Amostral para Série Combustível



Fonte: Software R

A Figura 9 destaca que a Função de Autocorrelação (FAC) e a Função de Autocorrelação Parcial (FACP) exibem um comportamento que reflete uma combinação dos padrões observados em modelos autoregressivos e médias móveis. Isso é evidenciado pelo decaimento gradual da FAC em direção a zero e o decaimento semelhante da FACP.

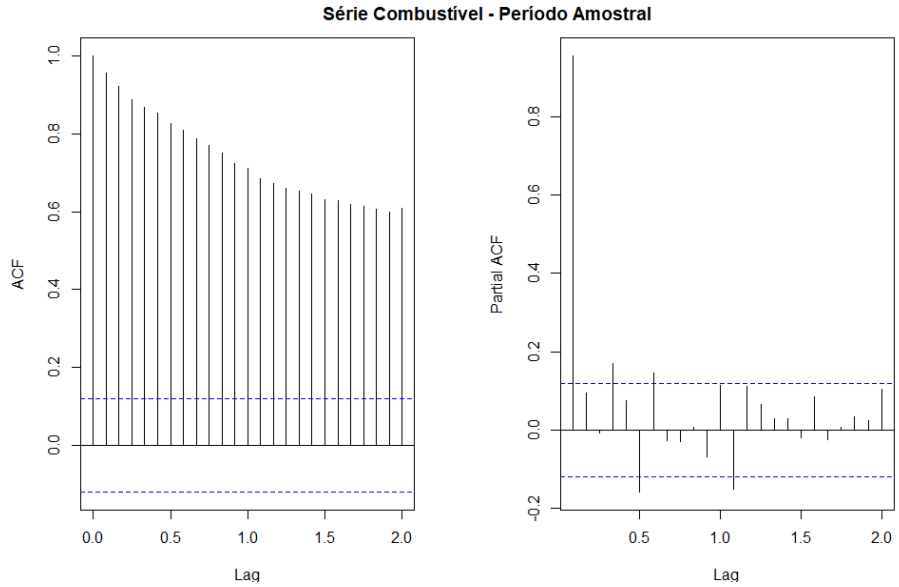
Como resultado, se conduziram vários testes de modelos ARMA, e se observou que o ARIMA (2,1,1)(0,1,1) S=12 se alinha de maneira mais consistente com as características da série.

Utilizando o procedimento de identificação automática de modelos fornecido pelo software R, denominado 'autoarima', obteve-se como resultado um modelo SARIMA (0,1,1) (0,1,1) [12].

Essa escolha do modelo pelo Software conforme a Tabela 1 é baseada em critérios estatísticos, como o critério de informação de Akaike (AIC) ou o critério de informação bayesiano (BIC). Esses critérios buscam encontrar o modelo que melhor se ajusta ao conjunto dos dados, considerando a complexidade do modelo

Figura 9 - Função de autocorrelação e função de autocorrelação parcial da série combustível

no período amostral



Fonte: Software R

Para se estimar os parâmetros do modelo, emprega-se a função "arima()" dentro do software R, específica para processos ARMA. Esta função possibilita a utilização de três métodos de estimação, sendo que se optará exclusivamente pelo CSS-ML (Conditional Sum of Squares - Maximum Likelihood). Os parâmetros do modelo de combustível no período amostral foram obtidos na Tabela 1.

A seguir serão apresentadas as equações dos modelos SARIMAS:

- SARIMA (0,1,1)(0,1,1)[12]

$$(1-B)(1-B^s)Y_t = (1-\theta)(1-\Theta B^s)a_t$$

$$Y_t = Y_{t-1} + Y_{t-s} - Y_{t-(s+)} = a_t - \theta a_{t-1} - \Theta a_{t-s} + \theta\Theta a_{t-(s+1)}$$

- SARIMA (2,1,1)(0,1,1)[12]

$$(1-\phi_1 B - \phi_2 B^2)(1-B)(1-B^s)Y_t = (1-\theta)(1-\Theta B^s)a_t$$

Portanto para a toma de decisão será escolhido o modelo com o valor mais baixo de AIC, o qual é considerado o mais adequado juntamente com a análise das outros critérios de avaliação, indicando um melhor equilíbrio entre ajuste e complexidade conforme a Tabela 1.

Identificando um modelo e estimado os respectivos parâmetros passa-se à fase de avaliação do diagnóstico. Nesta fase começa-se por analisar a qualidade estatística do modelo estimado. Portanto, um modelo que não satisfaça os seus pressupostos deve ser rejeitado, e então se repetir o ciclo.

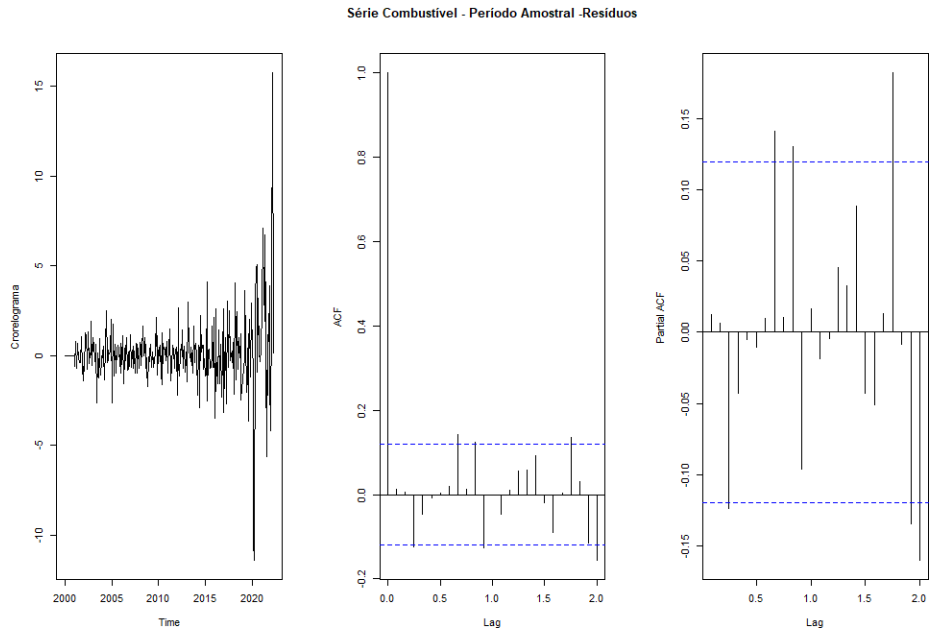
Identificação → Estimação → Avaliação do diagnóstico

Até se encontrar um modelo satisfatório para descrever a sucessão em causa. No caso de existirem diversos modelos de boa qualidade à luz dos seus pressupostos, põe-se o problema de escolha do melhor modelo. A resolução deste problema pode ser efetuada através de alguns critérios, de entre os quais se salienta o critério AIC.

Tendo em conta o princípio de parcimônia, foi realizado o teste de significância do modelo da série combustível através da função coefteste() e todos os parâmetros apresentam significância estatística nos dois modelos conforme a Tabela 1. Portanto se mantem os modelo com os parâmetros que foram achados.

Conforme a Figura 10 foram geradas representações gráficas dos cronogramas, da função de autocorrelação (FAC) e da função de autocorrelação parcial (FACP). No entanto, é importante observar que na FACP há (5) cinco lags fora das bandas de confiança, indicando que não estamos exatamente diante de um ruído branco específico, embora a semelhança seja notável.

Figura 10 - Correlograma, função de autocorrelação e função de correlação parcial dos resíduos da série combustível no período amostral.



Fonte: Software R

Procede-se à aplicação de testes. Devido ao número de resíduos serem superior a 50, precisamente 269 observações, é empregado o teste de Kolmogorov-Smirnov. Os resultados indicam que os resíduos resultantes do ajuste não seguem uma distribuição de ruído branco, pois o p-value é igual a $2.2e^{-16}$ para a série do Modelo (1) e $2.227e^{-09}$ para a serie do Modelo 2, conforme evidenciado na Tabela 1.

Portanto, não há conformidade com uma distribuição normal.

O teste de Ljung-Box ajuda a verificar se os resíduos de um modelo de séries temporais são independentes e não apresentam padrões significativos de autocorrelação, o que é crucial para a validade do modelo. Esse teste, ao utilizar a estatística qui-quadrado χ^2, avalia se há autocorrelação nos resíduos em diferentes defasagens, sendo essencial para garantir a confiabilidade das análises e previsões realizadas. A aplicação deste teste evidencia que se rejeite a independência dos dados conforme a Tabela 1, pois o seu p-valeu resultou em 0.003895 para o Modelo (1) e 0.06388 para o Modelo (2).

Tabela 1 - Estimação e testes de significância dos modelos e pressupostos de normalidade e independência dos resíduos da série combustível.

Estatísticas	Modelo (1) SARIMA (0,1,1)(0,1,1)		Modelo (2) SARIMA (2,1,1)(0,1,1)	
Teste de Significância dos Coeficientes				
Parâmetro	Coef.	p-valor(>|z|)	Coef.	p-valor(>|z|)
$\emptyset_1$	-	-	0.8986	$7.773e^{-12}$
$\emptyset_2$	-	-	-0.3554	$2.475e^{-08}$
θ_1	0.3121	$9.538e^{-08}$	-0.5724	$1.262e^{-05}$
Θ_1	0.585	$2.2e^{16}$	0.0820	$2.2e^{-16}$
Estatística dos Modelos				
MSE	720.24		612.87	
AIC	1122.18		1228.56	
BIC	1127.36		1246.52	
Estatística dos Resíduos				
Média	0.1103		0.1531	
Desvio Padrão	2.0570		2.2860	
Assimetria	0.8329		0.5498	
Curtose	17.5028		11.6215	
Teste dos Pressupostos Estatísticos dos Resíduos				
Normalidade	$2.2e^{-16}$		$2.227e^{-09}$	
Independência	0.003895		0.06388	

Fonte: elaboração própria.

Conforme a Tabela 1, o teste de significância para os dois modelos foram significativos, no entanto é perceptível que o Modelo (1) apresenta características

como simplicidade, facilidade de ajuste e interpretação mais adequadas para sua análise devido ao número de parâmetros que tem em comparação ao Modelo (2) escolhido através do procedimento de identificação automática do software R.

Porém, para corroborar essa informação se observará algumas estatísticas, tais como: MSE – Erro Médio Quadrático, o qual reflete a média dos quadrados dos erros entre as previsões e as observações. O critério de comparação é quanto menor melhor o desempenho do modelo em termos de previsão, visto isso, se observa que o Modelo (2) apresenta um desempenho ligeiramente menor ao Modelo (1). Além disso, têm-se outras medidas de qualidade como: O Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), os quais penalizam modelos mais complexos, tendo como critério de avaliação que um valor menor indica um modelo mais preferível, demonstrando uma preferência para o Modelo (1).

A seguir serão avaliadas as estatísticas dos resíduos que conforme evidenciado na Tabela 2, ambos os modelos apresentam médias de resíduos próximas de zero, indicando que são relativamente não enviesados, também apresentam desvios padrão semelhantes, indicando dispersão muitos próximas aos resíduos. Quanto à assimetria, ambos os modelos apresentaram assimetria positiva, sugerindo caudas mais longas na distribuição dos resíduos. Além disso, a Curtose é alta, sugerindo caudas pesadas e propensas a valores extremos, neste quesito o Modelo (1) tem maior susceptibilidade a este critério de avaliação, o que pode ser um indicador do contexto específico da série combustível no período avaliado, por exemplo, exibir caudas pesadas devido a eventos raros como a pandemia do COVID-19 ou a necessidade de avaliar o modelo, considerando transformações nos dados ou explorar a inclusão de componentes adicionais no modelo.

Por último como explicado anteriormente nenhum dos modelos alcançou satisfazer os Teste dos Pressupostos Estatísticos dos Resíduos de Normalidade e Independência, o que em certa medida compromete o diagnóstico sobre a boa qualidade das previsões que possam surgir a partir do modelo escolhido. Isso significa que há problema de autocorrelação nos resíduos indicando que os modelos provavelmente não capturaram completamente a estrutura temporal dos dados, e pode haver informações não exploradas que poderiam melhorar as previsões.

Tabela 2 - Resumo das Estatísticas dos Erros de Previsão.

Métricas de Avaliação de Modelos	Modelo 1 SARIMA (0,1,1)(0,1,1)	Modelo 2 SARIMA (2,1,1)(0,1,1)
ME	0.1103	0.1531
RMSE	2.0561	2.2869
MAE	1.1949	1.4089
MAPE	0.0500	0.1667
MPE	2.3969	2.9154
MASE	0.2563	0.3022

Fonte: elaboração própria.

Interpretando os resultados apresentados na Tabela 2, que avalia o desempenho de ambos os modelos, se começará com a análise do Erro Médio Absoluto (MAE). O MAE utiliza o módulo de cada erro, mitigando a subestimação e tornando-se menos sensível a valores extremamente discrepantes (outliers). Essa abordagem é crucial, uma vez que impede que pontos extremos tenham um impacto desproporcional nos resultados.

Ao considerar o MAE como métrica de avaliação, buscamos evitar a tendência de avaliar a precisão de um modelo apenas pela magnitude do desvio em relação ao valor real, sem fazer distinção entre valores acima ou abaixo da verdadeira medida. Nesse contexto, é notável que o Modelo (1) demonstra um desempenho superior em relação a esta métrica quando comparado ao Modelo (2).

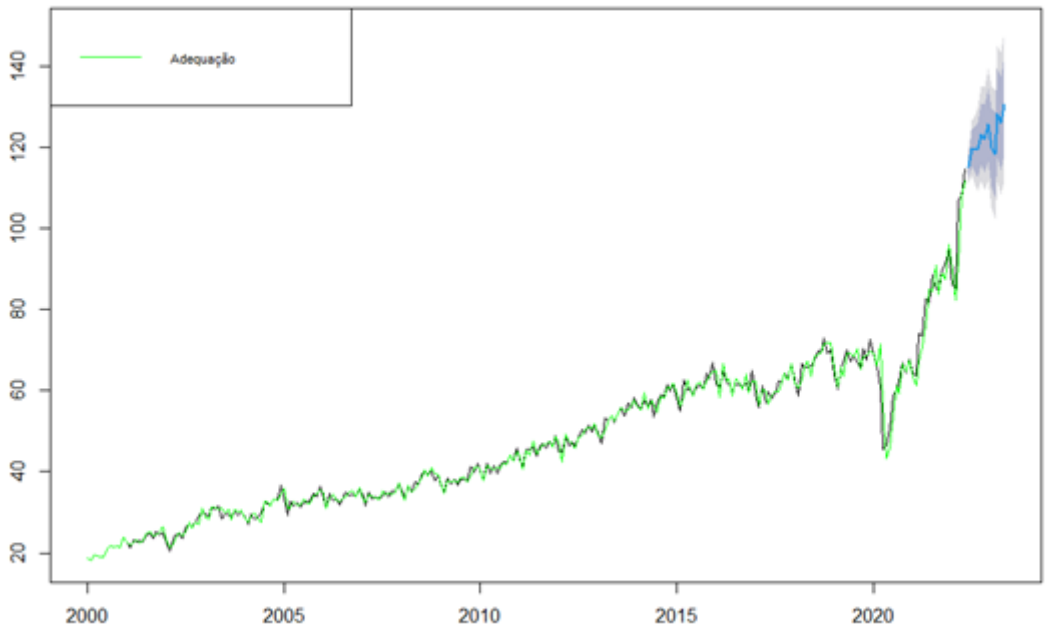
Prosseguindo com a análise, abordaremos agora o Erro Quadrático Médio (RMSE), uma métrica que avalia a precisão dos modelos, atribuindo maior peso aos erros mais significativos. Ao calcular o RMSE, cada erro é individualmente elevado ao quadrado, e a média desses erros quadráticos é então calculada. Essa abordagem favorece a detecção e penalização de desvios mais expressivos.

É importante ressaltar que o RMSE é particularmente sensível a valores extremos, dado que o processo de elevar os erros ao quadrado amplifica a contribuição desses pontos discrepantes. Nesse contexto, observa-se que o Modelo (1) apresenta um desempenho superior, evidenciado pela presença de erros menores em comparação com o Modelo (2).

A partir desse critério de seleção conforme as Tabelas 1 e 2 se utilizará o

Modelo (1) para realizar as previsões a partir do período amostral, assim, para avaliar os resultados, primeiramente se diagnosticará a eficácia do Modelo (1) SARIMA (0,1,1)(0,1,1) [12] ajustado a 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 Janeiro de 2000 a Maio de 2022, com os valores calculados pelos parâmetros do modelo conforme a Figura 11 que está com um linha em verde sobrepondo ao gráfico original.

Figura 11 - Série Treinamento versus Valores Ajustado pelo Modelo SARIMA(0,1,1)(0,1,1) à variação no período amostral



Fonte: Software R

Como se pode observar a série ajustada acompanha bem a série gerada, do ponto de vista da representação gráfica, inclusive, com os picos e vales causados pela sazonalidade identificada na maioria dos períodos, o modelo parece ser adequado, apesar dele não cumprir os pressupostos estatísticos de Normalidade e Independência.

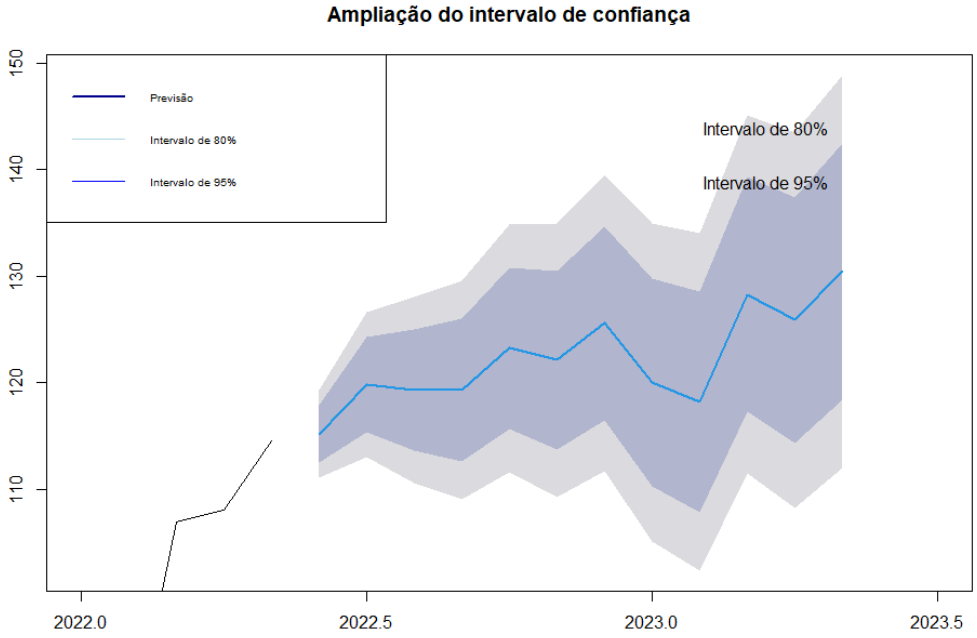
Além disso, se construiu o gráfico das previsões 12 passos à frente, (Lim. Inf. das Prev., Lim. Sup. das Prev., e as obs. No período de validação). O modelo não consegue acompanhar a variabilidade da série de interesse no período de validação conforme o Figura 8.

Observa-se que a maioria das previsões se manteve dentro dos intervalos de

confiança de 80% e 95%. Mesmo quando as previsões ficaram abaixo de 80%, ainda podemos avaliar o desempenho como aceitável, especialmente considerando que se tratava de projeções para 12 passos à frente durante um evento inesperado: a pandemia de COVID-19, que teve início no primeiro trimestre de 2020.

É importante destacar que o modelo enfrentou dificuldades para captar esse período de sazonalidade atípica, conforme evidenciado na Figura XX (anterior). Contudo, em períodos sem a influência desse evento, o modelo demonstrou comportamento adequado ao descrever e acompanhar a tendência e volatilidade de maneira mais "normal".

Figura 12 - Gráfico das previsões 12 passos à frente



Fonte: Software R

A queda no desempenho das previsões pode ser explicada pelo histórico da chegada da COVID-19 no Brasil, gerando incertezas relacionadas à crise sanitária. Isso resultou no fechamento de comércios em todo o país e, consequentemente, em uma abrupta redução da atividade econômica de 2020 a 2022. Além disso, durante o período de previsão a partir de maio de 2022, o Brasil enfrentava uma situação inflacionária devido à desaceleração da economia global.

No contexto da pandemia, houve um aumento significativo nos gastos públicos para apoiar as pessoas mais vulneráveis, por meio de incentivos para ficarem em casa, como o programa "Auxílio Brasil", e outros investimentos em saúde, embora tenha ajudado a conter as crescentes quedas refletidas na série não foram suficientes,

principalmente porque o nível de desemprego e consumo de produtos a nível geral tiveram baixas expressivas. Em resumo, observaram-se quedas drásticas devido ao fechamento compulsório de comércios, seguidas por uma retomada das atividades ao "normal" e um crescimento da demanda por produtos que não refletiam o processo natural de desenvolvimento da série temporal de combustíveis, mesmo com picos e vales foi o que proporcionou as maiores dificuldades para o modelo prever de maneira adequada esse “outliers” ocorrendo de maneira simultânea.

7 CONCLUSÃO

O modelo construído através da metodologia Box-Jenkins apresenta de maneira inicial ter cumprido seu propósito. Os valores previstos para a série de teste tiveram boa aderência aos dados reais, mesmo com a instabilidade econômica e social causada pelo evento extremo ocorrido a partir do ano de 2020 (a pandemia COV-19). No entanto, embora os dados previstos para 12 passos à frente não obtiveram um resultado tão satisfatório conforme discutido na seção anterior, acredita-se que há possibilidade de melhorar o desempenho através da inclusão de variáveis exógenas ao modelo ou proveito de outras variáveis de desempenho no âmbito do varejo como alimentos, vestimenta entre outras, com o intuito de apresentar alternativas que possam melhorar a capacidade preditiva como aplicação de uma regressão dinâmica.

O propósito deste trabalho consiste em avaliar o desempenho no processo de previsão entre os dois modelos analisados. Nesse contexto, observa-se que o Modelo (1) demonstra indicadores superiores em termos de precisão de previsão e sensibilidade a erros quando comparado ao Modelo (2). Essa constatação sugere que, para os critérios e métricas estabelecidos, o Modelo (1) emerge como a escolha mais eficaz e confiável na realização das predições propostas pelo estudo.

REFERÊNCIAS

Babu, C. N.; Reddy, B. E.. Predictive data mining on average global temperature using variants of ARIMA models. In: **IEEE International Conference On Advances In Engineering, Science and Management (ICAESM-2012)**, Tamil Nadu, p. 256-260, 2012.

Chen, C.; Hu, J.; Qiang, M.; Yi, Z.. Short-time traffic flow prediction with ARIMA-GARCH model. In: **2011 IEEE Intelligent Vehicles Symposium (IV)**, Baden-Bande, p.607-612, 2011.

Donselaar, K. H.; Gaur, V.; Woensel, T.; Broekmeulen, R. A. C. M.; Fransoo, J. C.. Ordering Behavior in Retail Stores and Implications for Automated Replenishment. In: **Management Science**, v.56, n.5, p. 766-784, 2010.

Fattah, J.; Ezzine, L.; Aman, Z.; Moussami, H.; Lachhab, A.. Forecasting of demand using ARIMA model. In**: International Journal of Engineering Bussiness Management**, v.10, p.9, 2018.

Fildes, R.; Shaohui, M.; Kolassa, S.. Retail forecasting: Research and practice. In: **International Journal of Forecasting**, v.38, n.4, p.1283-1318, 2022.

Ghafour, M. K.; Aljanabi, A. R. A.. The role of forecasting in preventing supply chain disruptions during the COVID-19 pandemic: a distributor-retailer perspective. In: **Operations Management Research**, v.16, n.2, p. 780-793, 2023.

Gilbert, K.. An ARIMA supply chain model. In: **Management Science**, v.51, n.2, p.305-310, 2005.

Gujarati, D. N.; Porter, D. C. Econometria básica-5. **Amgh Editora**, 2011.

Miotto, A. P.; Parente, J. G.. Retail evolution model in emerging markets: apparel store formats in Brazil. In**: International Journal of Retail & Distribution Management**, v.43, n.3, p. 242-260, 2015.

Morettin, P. A,; Toloi, C. M. C. Análise de Série Temporais. São Paulo, Edgar Bluncher, 2004.

Olsson, M.; Soder, L. Modeling real-time balancing power market prices using combined SARIMA and Markov processes. In: **IEEE Transactions on Power Systems**, Manchester, v.23, n.2, p. 443-450, 2008.

SEBRAE. Varejo no Brasil: cenário atual, futuro e oportunidades. Minas Gerais, agosto, 2023. Disponivel: https://sebraeplay.com.br/content/varejo-no-brasil-cenario-atual-futuro-e-oportunidades.

Wang, J.; Chong, W. K.; Lin, J.; Hedenstierna, C. P.. Retail Demand Forecasting Using Spatial-Temporal Gradient Boosting Methods. In: **Journal of Computer Information Systems**, p. 1-13, 2023.

Wang, J.; Liu, G. Q.; Liu, L.. A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry. In: **2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)**, Chengdu, p. 317-320, 2019.

World Bank**. Global Economic Prospects**. Washington, DC, 2021.