UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma análise temporal da influência do IPCA, taxa de desemprego e renda média no consumo de crédito da população brasileira

Arthur Lucena Silva

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uma análise temporal da influência do IPCA, taxa de desemprego e renda média no consumo de crédito da população brasileira

Arthur Lucena Silva

USP - São Carlos

2023

Arthur Lucena Silva

Uma análise temporal da influência do IPCA, taxa de desemprego e renda média no consumo de crédito da população brasileira

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Aprendizado de máquina

Orientador: Prof. Dr. Ricardo Araújo Rios

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S586a

Silva, Arthur Lucena
Uma análise temporal da influência do IPCA, taxa de desemprego e renda média no consumo de crédito da população brasileira / Arthur Lucena Silva; orientador Ricardo Araújo Rios. -- São Carlos, 2023.
55 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Séries temporais. 2. Ciência de dados. 3. Machine Learning. 4. Consumo de crédito. 5. Dados públicos. I. Rios, Ricardo Araújo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Aos meus 2 cunhados, Júnior e Vinicius, pela busca insaciável por conhecimento.

AGRADECIMENTOS

Às minhas irmãs Andrea Cristina Lucena Furlan, e Amanda Lucena Silva, pelo apoio, amor, e admiração incondicional.

Ao meu orientador Prof. Dr. Ricardo Araújo Rios, pelo compartilhamento de seu grande conhecimento, apoio e incentivo do inicio ao fim do trabalho.

Ao meu líder Jose Afonso Meirelles, por em momentos de tristeza e de dificuldade me apresentar novos caminhos a seguir e por me tornar mais resiliente.

À minha líder Monique Moya Guerreiro, por ter me creditado a oportunidade de conhecer o vasto universo de meios de pagamento e compartilhar sua visão crítica sobre a informação.

Ao meu time da Elo: Beatriz de Oliveira Telles, Erica de Souza Farjado, Felipe Eduardo Rodrigues, Gabriela Ayumi Ueda, Gustavo Paranhos Jurado, Leonardo Pedroso Dourado, Rodrigo Kendi Nakamura e Rodrigo Moreira Santos, por tornarem meu ambiente de trabalho tanto um lugar de excelência, como também um lugar de amizade, me despertando a força e a vontade de continuar estudando.

EPÍGRAFE

Para quem tem pensamento forte, o impossível é só questão de opinião.

Alexandre Magno Abrão (2013)

RESUMO

LUCENA, S. A. **Uma análise temporal da influência do IPCA, taxa de desemprego e renda média no consumo de crédito da população brasileira**. 2023. 70 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A crescente demanda por crédito da população brasileira é motivo de estudo em muitas empresas que lidam com meios de pagamento eletrônico. É vital para essas companhias poder elaborar um modelo que seja capaz de refletir algum poder preditivo sobre essa demanda para que estrategicamente estas organizações possam se planejar para o que está por vir, ou adotar medidas de inovação, dentre outros. Procurar entender como outras variáveis podem contribuir na predição de crédito é o objetivo deste trabalho. Estudiosos do setor econômico enxergam uma correlação entre o IPCA e o desemprego como influenciadores sobre a demanda de crédito, principalmente sobre a população brasileira de baixa renda. Uma vez que uma certa pessoa alcança um padrão de vida, ela irá procurar mantê-lo, mesmo que isso signifique obter mais crédito. Portanto, é necessário compreender como variáveis econométricas como IPCA, taxa de desemprego e renda média, podem influir sobre o comportamento preditivo do modelo. O impacto da COVID-19 trouxe dificuldade em prever crédito para os anos posteriores a 2020, refletindo aumento do resíduo aleatório e mudança nos picos de gasto observados na decomposição da série. O que levantou a questão sobre a mudança de comportamento da população brasileira após a pandemia, contudo, através de técnicas de aprendizado de máquina foi possível investigar e preparar a base de dados para conjecturar um modelo baseado em série temporal univariada com a série de crédito que apresentasse alguma utilidade preditiva, servindo como referência para a comparação dos modelos multivariáveis. Para a transformação da série não estacionária em estacionária foi aplicado uma diferenciação logarítmica, em seguida os modelos SARIMA, Prophet e Holt-Winter foram experimentados via pacote da biblioteca do Facebook denominada Kats. Os dados enviesados pela pandemia foram substituídos em função da predição do modelo univariado Prophet, utilizando como entrada os dados pré pandemia. Posteriormente com os dados obtidos a partir do melhor modelo testado, foi possível realizar os experimentos uni e multivariados. Os modelos multivariados não foram significativamente melhores que o modelo univariado, entretanto, foi possível identificar uma melhora progressiva na construção de um modelo com crédito e renda média, seguido de um modelo com crédito, renda média e IPCA, respectivamente. Este último atingiu o melhor resultado experimental dentre todos, comprovando que a utilização de variáveis econométricas pode trazer algum benefício para o trabalho, desde que sejam analisadas de forma cruzada.

Palavras-chave: Aprendizado de Máquina; COVID-19; IPCA; Resíduo Aleatório; Série Temporal.

ABSTRACT

LUCENA, S. A. **A temporal analysis of the influence of IPCA, unemployment rate and average income on credit consumption of the Brazilian population**. 2023. 70 f. Course completion work (MBA in Artificial Intelligence and Big Data) - Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2023.

The growing demand of the brazilian population for credit is a reason of study in many companies that deal with electronic ways of payment. It is very valuable for these companies to be able to develop a model wich can reflect some predictive power about this demand. That way these organizations can strategically plan for what is to come, or adopt innovation measures, among others. Seeking to understand how other variables can contribute to the prediction of credit is the objective of this work. Economic scholars see a correlation between the IPCA and unemployment as influencers on credit demand, especially on the brazilian low-income population. Once a certain person achieves a standard of living, they will seek to maintain it, even if it means obtaining more credit. Therefore, it is necessary to understand how econometric variables such as IPCA, unemployment rate and average income, can influence the predictive behavior of the model. The impact of COVID-19 brought difficulty in predicting credit for the years after 2020, reflecting an increase in the random residual and a change in the spending peaks observed in the decomposition of the series. This raised the question about the change in behavior of the brazilian population after the pandemic, however, through machine learning techniques it was possible to investigate and prepare the database to conjecture a model based on a univariate time series with the credit series that presents a positive predictive effect, serving as a reference for the comparison of multivariable models. To transform the non-stationary series into a stationary one, logarithmic differentiation was applied, then the SARIMA, Prophet and Holt-Winter models were tested via the Facebook library package called Kats. Data biased by the pandemic was replaced as a function of the prediction of the univariate Prophet model, using pre-pandemic data as the input. Subsequently, with the data obtained from the model, it was possible to proceed the univariate and multivariate experiments. The multivariable models were not significantly better than the univariable models, however, it was possible to identify a progressive improvement in the construction of a model with credit and average income, followed by a model with credit, average income and IPCA, respectively. The latter achieved the best experimental result among all, proving that the use of econometric variables can bring some benefit to the work, as long as they are cross-analyzed.

Palavras-chave: Machine Learning; COVID-19; IPCA; Random Residual; Temporal Series.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABECS	–	Associação Brasileira das Empresas de Cartões de Crédito
IBGE	–	Instituto Brasileiro de Geografia e Estatística
IPCA	–	Índice Nacional de Preço ao Consumidor amplo
IA	–	Inteligência Artificial
KATS	–	Kit to Analyze Time Series
CRISP-DM	–	Cross Industry Standard Process for Data Mining

SUMÁRIO

1. INTRODUÇÃO

Segundo levantamento feito pelo Serasa em todo o Brasil, cerca de 70% das pessoas que utilizam cartão de crédito, possuem 3 ou mais cartões. Comenta-se também que a inflação em alta por vários meses, faz com que muitas famílias optem por mais cartões para manter o padrão de consumo (P. Guimarães, 2022).

A ABECS (Associação Brasileira das Empresas de Cartões de Crédito e Serviços), setor oficial responsável pelos meios eletrônicos de pagamento no Brasil, aponta em seu relatório trimestral de 2022 que o uso do cartão de crédito para pagamentos aumentou mais de 40% no primeiro trimestre de 2022 quando comparado ao mesmo período em 2021. Em acompanhamento à modalidade crédito, o uso de cartões de débito e pré-pago também cresceu no período. As principais compras onde a população considera o uso do cartão de crédito mais importante são respectivamente em supermercado e alimentação (17%), farmácia (15%) e eletrodomésticos (14%).

Outro conceito importante que influencia economicamente as famílias brasileiras é o índice oficial da inflação no país, medido pelo IPCA (Índice Nacional de Preços ao Consumidor Amplo). Como o próprio nome sugere, ele indica a variação de preços de um conjunto de produtos e serviços para o consumidor final, refletindo o consumo pessoal das famílias. O indicador é calculado mensalmente pelo IBGE (Instituto Brasileiro de Geografia e Estatística) nas principais regiões metropolitanas do país. Apesar de não ser calculado em todo o país, possui abrangência nacional.

O IPCA reflete a inflação, i.e., quando o número de pessoas interessadas em um determinado produto aumenta rapidamente, fica difícil garantir o fornecimento para todos. Nesse caso, entende-se que a demanda superou a oferta, provocando o aumento no preço do item buscado para que o número de pessoas interessadas no mesmo item diminua, este evento é denominado inflação. A recíproca também pode gerar inflação, quando os consumidores possuem maior poder de crédito, a demanda geral passa a aumentar, tornando difícil o fornecimento para todos, provocando o mesmo evento.

Outro fator que reflete a inflação é a taxa de desemprego. Segundo a Folha de São Paulo em artigo publicado por Canzian, F. (2021) o aumento da inflação sobre os produtos alimentícios em conjunto com o desemprego em alta, provoca uma queda aguda do poder aquisitivo dos mais pobres, com agravantes de fome e miséria. O índice de desemprego, assim como o IPCA, também é calculado e disponibilizado pelo IBGE.

É possível identificar através das referências acima que há uma possível correlação entre o comportamento da população brasileira que faz uso de cartões de crédito, em função da variação do IPCA e do desemprego. A partir desta hipótese, deseja-se investigar se há de fato uma associação presente nos dados que possa ser incorporada por uma Inteligência Artificial (IA), contribuindo para uma predição eficiente sobre o volume de crédito gasto.

1.1. Justificativa

Esta pesquisa tem potencial estratégico determinante no planejamento comercial de empresas no ramo de meios de pagamento eletrônicos. O modelo de predição desenvolvido, caso se apresente eficiente, pode influenciar na redução de custos de produção de cartões, redução de rotatividade de clientes, otimização de recursos e definição de metas mais adequadas à realidade conforme os parâmetros investigados. Caso contrário, pode servir como argumento empírico para seguir em outras direções, evitando investimentos desnecessários e perda de tempo com ideias não confirmadas perante os dados.

1.2. Objetivo

Construir um modelo preditivo útil capaz de estimar a volumetria de gastos em cartões de crédito em determinados meses do ano.

1.2.1. Objetivos específicos

- Criar e testar modelos uni e multivariados para predição de crédito baseado em análise temporal;
- Identificar se o IPCA aumenta a eficácia do modelo;
- Identificar se a taxa de desemprego aumenta a eficácia do modelo;
- Identificar se a renda média formal aumenta a eficácia do modelo;
- Identificar se o IPCA em conjunto com a taxa de desemprego e, ou, com a renda formal média, aumenta a eficácia do modelo.

2. REVISÃO BIBLIOGRÁFICA

2.1. Classificação do problema

O presente problema é visto como uma tarefa supervisionada, cujo objetivo principal é modelar os componentes sazonais, estocásticos e de tendência de séries temporais. Dentre tais componentes, o foco principal deste trabalho será modelar a sazonalidade, considerando a hipótese de que em determinadas épocas do ano devido à presença de eventos de natureza comercial ou cultural, há um aumento no volume de crédito gasto pela população, que mobiliza empresas a se prepararem para a demanda que surge nestas épocas específicas, como, por exemplo, a páscoa que ocorre em abril e o natal que acontece em dezembro.

Dados de comportamento humano, geralmente contém alguma forma de sazonalidade, mesmo com diversos ciclos (e.g., um padrão de horário, semanal, verão-inverno). As seções a seguir apresentam mais informações sobre componentes que influenciam o comportamento de séries temporais e ferramentas de modelagem de suas observações.

2.2. Séries temporais

A análise de séries temporais visa resumir e extrair informações estatísticas de séries temporais, no intuito de diagnosticar comportamentos passados e predizer observações futuras (A. Nielsen, 2021).

O estudo de séries envolve a introdução de suposições simplificadoras que conduzem a análise, como identificar se suas observações são estacionárias. Uma série é denominada estacionária quando esta se desenvolve ao longo do tempo em torno de uma média constante, refletindo alguma forma de equilíbrio ou estabilidade. Uma série não estacionária apresenta uma tendência em função do tempo, a qual pode ser positiva ou negativa. A tendência não é necessariamente linear, podendo ser explosiva como o crescimento de uma colônia de bactérias ou um objeto acelerando em queda livre (Pedro A. Morettin e Clélia M. C. Toloi, 2006).

2.3. Modelos estatísticos

Existem diversos algoritmos com foco em análise temporal, alguns dos mais conhecidos devido sua boa performance são os autorregressivos, baseados em média móvel, e os aditivos

que recebem parâmetros que representam fatores que vão influenciar na predição. Esses, e outros tipos de modelos disponíveis na KATS (*Kit to Analyse Time Series*) foram experimentados na pesquisa, e cada um possui uma mecânica de predição específica que pode corresponder melhor ao sistema à depender das características temporais do conjunto de dados.

2.4. Características da série temporal

De acordo com Box & Jenkins (1970), informações de sazonalidade e tendência precisam ser estimadas para possibilitar uma compreensão adequada dos modelos e um ajuste otimizado dos parâmetros para obtenção de melhores resultados. A Figura 1 ilustra uma série temporal criada a partir de dados dos passageiros das companhias aéreas dos EUA coletados entre 1949 à 1961, que é um exemplo claro de uma série temporal não estacionária por exemplificar um comportamento crescente e padrões sazonais. Também é possível notar evidências de sazonalidade refletindo intrinsecamente um processo não estacionário, ou seja, tendencioso.

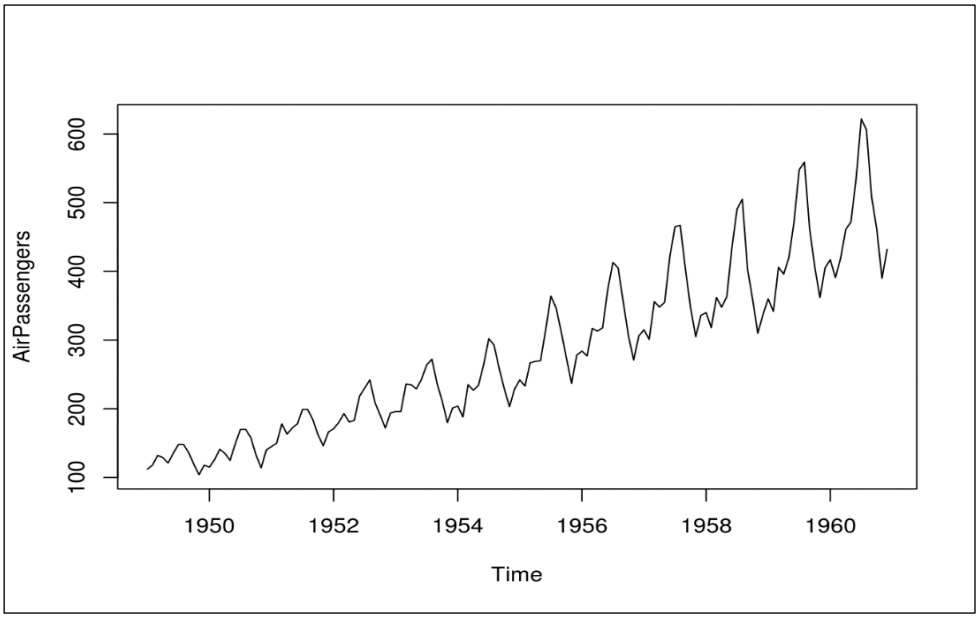

Figura 1 – Crescimento de passageiros das companhias aéreas dos EUA entre 1949 e 1961

Fonte: Aileen Nielsen (2021).

A Figura 2 ilustra os componentes de tendência e sazonalidade obtidos a partir da decomposição da série citada anteriormente, além do erro residual. Nas seções a seguir, esses componentes são descritos com mais detalhes.

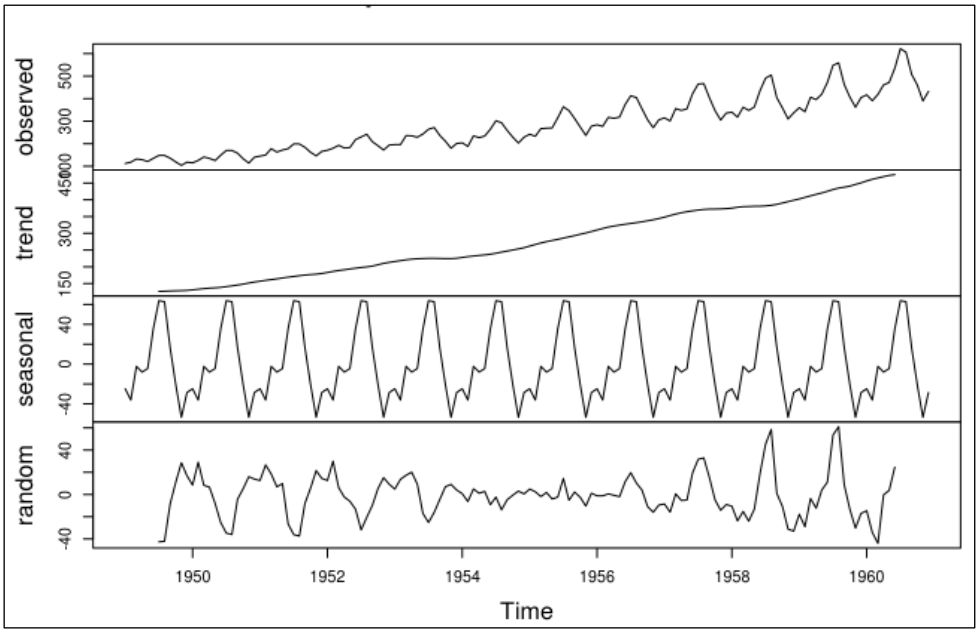

Figura 2 – Decomposição da série de passageiros das companhias aéreas dos EUA entre 1949 e 1961

Fonte: Autoral (2023).

2.4.1. Tendência

Representa o padrão, e a velocidade, de crescimento, decrescimento, ou estabilidade, da série temporal em um determinado período de tempo. Na Figura 2, a varável *trend* representa este padrão de comportamento através de uma reta crescente em função do tempo. Geralmente trabalha-se com tendência constante, linear ou quadrática, conforme ilustrado na Figura 3.

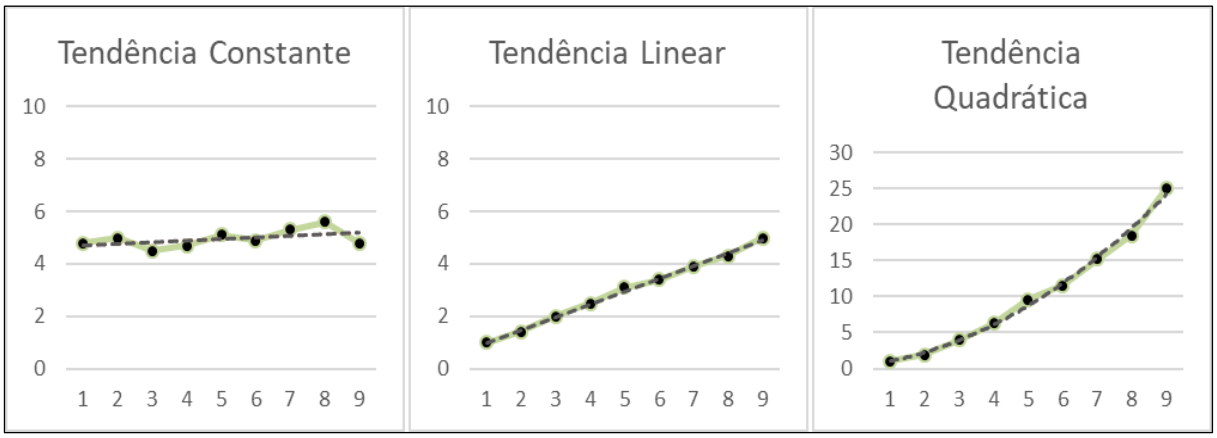

Figura 3 – Tendências mais comuns em séries temporais

Fonte: Autoral (2023).

2.4.2. Sazonalidade

Trata-se do padrão de comportamento da série temporal que se repete em épocas específicas. Além da sazonalidade, também é válido citar que existem comportamentos cíclicos e, conforme descrito por Aileen Nielsen, no livro análise prática de séries temporais, tais características possuem diferenças:

> As séries temporais sazonais são séries temporais em que os comportamentos se repetem durante um período fixo. É possível ter várias periodicidades refletindo diferentes ritmos de sazonalidade, como a sazonalidade diária de 24 horas versus as estações do calendário de 12 meses, ambas as quais apresentam fortes características na maioria das séries temporais relacionadas ao comportamento humano.
> As séries temporais cíclicas também apresentam comportamentos recorrentes, mas elas têm um período variável. Um exemplo comum é um ciclo de negócios, como os ciclos de alta e baixa do mercado de ações, que têm uma duração incerta. Da mesma forma, os vulcões apresentam comportamentos cíclicos, mas não sazonais. Conhecemos os períodos aproximados de erupção, mas eles não são precisos e variam com o tempo (Aileen Nielsen, 2021, p.60).

2.4.3. Resíduo (ruído) aleatório

Outro componente importante na análise de séries temporais é a aleatoriedade. Após a remoção da tendência e da sazonalidade, conforme observado na Figura 2, a série resultante é caracterizada pela presença de ruído. O ruído é uma variável cujo padrão não pode ser precisamente modelado, representando a imprevisibilidade do sistema que produziu a série temporal. Quanto maior o resíduo, em geral, mais complexo torna-se a predição da série analisada. Por outro lado, desde que não seja erro de medição, o ruído pode representar uma informação valiosa quando representar picos ou degraus na série conforme a Figura 4, indicando a presença de eventos externos ao sistema estudado.

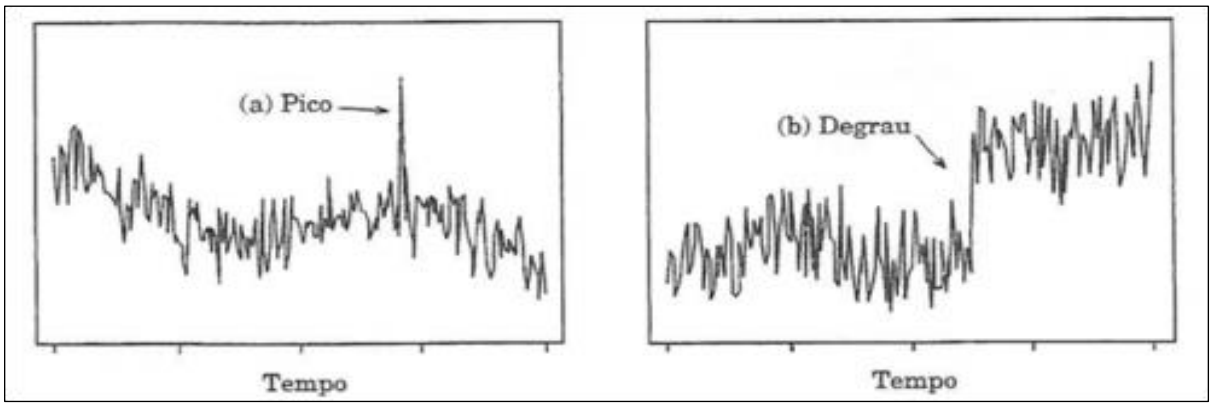

Figura 4 – Picos e degraus podem atrapalhar no ajuste do modelo, mas também informam sobre forças importantes que atuam no ambiente

Fonte: Wild & Seber (2004).

2.5. Características combinadas

Para demonstrar o impacto causado pelas características previamente citadas no âmbito de uma série temporal, a Figura 5 representa graficamente alguns exemplos que combinam os componentes de um sistema temporal de forma isolada.

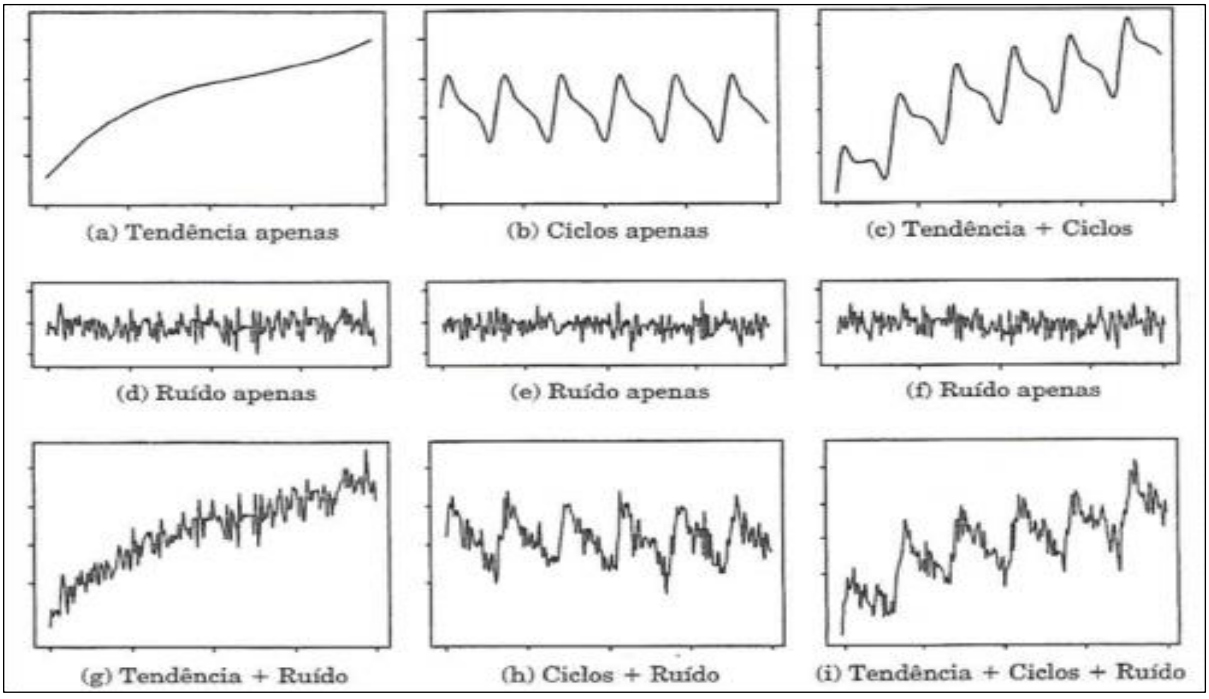

Figura 5 – Combinações de tendência, sazonalidade e ruído

Fonte: Wild & Seber (2004).

2.6. Método investigativo de mineração de dados

A metodologia para realizar um experimento em ciência de dados envolve uma variedade de conceitos tecnológicos, que, através de um processo bem definido, permite estruturar o problema de forma consistente tornando possível reproduzir todo o processo. Este processo é geralmente conhecido como mineração de dados, e um guia útil para conduzir tal procedimento é disponibilizado pela CRISP-DM (*Cross Industry Standard Process for Data Mining*). A Figura 6 ilustra o processo de mineração.

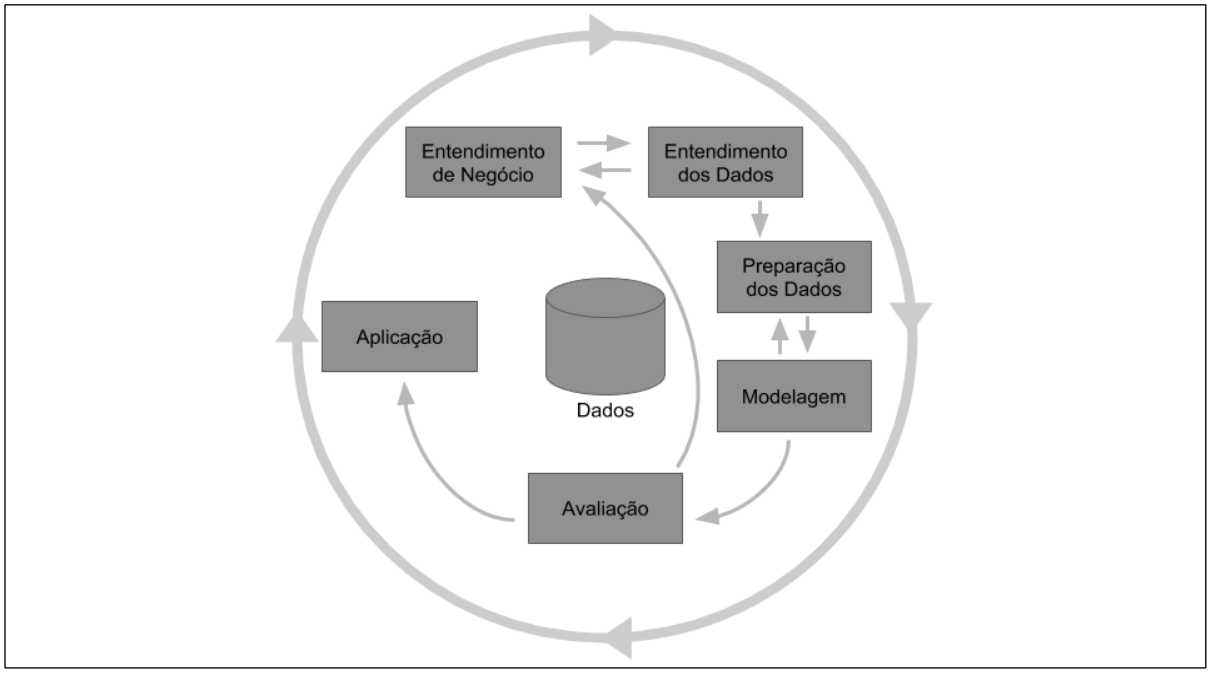

Figura 6 – Processo de mineração de dados CRISP-DM

Fonte: Karina Moura (2019).

O diagrama acima demonstra que é normal revisitar etapas do processo. Na verdade, esse é o fundamento da mineração de dados. Segundo F. Provost e T. Fawcett (2016), a cada volta dada no processo a equipe de ciência de dados tende a saber muito mais sobre o problema. Considerando essa metodologia como referência, foi possível organizar as seguintes etapas deste estudo:

1. **Compreensão do negócio:** O problema a ser resolvido deve ser devidamente entendido e classificado em termos de ciência de dados, com um nível suficiente de detalhes para que o caso torne-se específico, e para que o objetivo a ser atingido seja claro, e que o

seu impacto no negócio esteja evidente. Deve-se refletir a necessidade real de negócio e caso necessário decompor o problema em tarefas de mineração de dados. Deve-se refletir a necessidade real de negócio e caso necessário decompor o problema em tarefas de mineração de dados.

2. **Compreensão dos dados:** Matéria prima para a solução do problema, é importante entender os pontos fortes e as limitações dos dados porque raramente há uma correspondência exata com o problema estruturado. Dados históricos, muitas vezes, são recolhidos para fins não relacionados com o problema de negócio, isso pode fazer com que seja necessário a incorporação de mais dados no projeto, de diferentes tipos, para elaborar uma solução adequada e eficiente. Não é possível converter uma série de uma periodicidade baixa em uma série de periodicidade alta, por exemplo, semanal para diária ou diária para minutos, portanto, é fundamental compreender se o campo de tempo observado satisfaz a condição para propor uma solução adequada para o problema de negócio proposto. Para isso deve-se checar a disponibilidade das informações, e caso tais dados não existam, deve-se avaliar a possibilidade de obtenção das informações necessárias através de pesquisas, técnicas computacionais, ou demais meios alternativos.

3. **Preparação dos dados**: As tecnologias analíticas dentro do âmbito de ciência de dados são poderosas, mas impõem determinados requisitos sobre os dados. Se tratando de dados temporais é essencial traduzir corretamente o campo temporal para um formato que seja legível para a IA, e também garantir que os dados estejam íntegros e consistentes. Para tal, técnicas de reposição ou substituição de dados podem ser necessárias em caso, por exemplo, de dados ausentes ou discrepantes. Algumas outras tratativas podem ser necessárias conforme o conhecimento sobre o problema evolui.

4. **Modelagem**: A modelagem é algum tipo de modelo ou padrão que captura regularidade nos dados. Esta é a principal etapa onde as técnicas de mineração de dados são aplicadas.

5. **Avaliação:** Realizar o cruzamento dos resultados obtidos e determinar se as hipóteses sobre o problema foram verdadeiras, ou não, analisando os resultados obtidos de forma

quantitativa e qualitativa. Igualmente importante, a fase de avaliação também serve para ajudar a garantir que o modelo satisfaça os objetivos originais de negócio.

6. **Implementação:** Os resultados do modelo são postos a prova no mundo real, e são monitorados com rigor a fim de constatar algum retorno sobre o investimento.

3. MATERIAIS E MÉTODOS

3.1. Materiais

Os materiais utilizados na pesquisa resumem-se à interface adotada para escrita e disponibilização, dos códigos e dados necessários para realizar a análise, ambos detalhados conforme a seguir:

- **GitHub[1]:** Plataforma online pertencente ao autor onde disponibiliza-se a base de dados utilizada, o script elaborado para esta pesquisa, tanto na tratativa das informações, como também na análise e aplicação das técnicas utilizadas;
- **Anaconda (Jupyter Notebook):** Aplicação de código aberto voltado para desenvolvimento integrado à linguagem de programação Python, com foco em negócios;
- **Biblioteca para experimento:** Em concordância com Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72 (1), 37-45, A maioria dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias, o que faz com que seja necessário transformar os dados originais em caso de não haver estabilidade no sistema. Contudo, existem tecnologias que fazem este preparo, tornando o processo mais simplificado. Os pesquisadores do Facebook desenvolveram uma biblioteca chamada KATS, que conforme Facebook (2021), Kats - a toolkit to analyze time series data (v. 0.2.0). Disponível em: https://pypi.org/project/kats/ (Acessado em: 14 de Março de 2023), trata-se de uma IA generalizadora de estruturas de dados voltada para análise de séries temporais, que possibilita ao analista não despender parte do tempo da análise preparando os dados para torná-los aptos a serem incorporados por determinados modelos, além de permitir a identificação de *outliers* temporais, e análises uni e multivariadas, e por fim experimentar diversos tipos de modelos, facilitando e diversificando os experimentos;
- **Dados públicos da ABECS:** Série histórica de registros volumétricos de cartões de crédito, débito e pré-pago;
- **Dados públicos do IBGE:** Série histórica da variação do IPCA e também da taxa de desemprego e renda média.

[1] O GitHub é como um repositório onde é possível registrar base de dados, código, resultado e versões de um projeto. O link a seguir direciona o leitor para a o repositório onde estão armazenados os conteúdos utilizados para a análise deste projeto de pesquisa:
<https://github.com/tutalucena/Temporal-Series-Analyse-of-Credit-of-Brazillian-People>

3.2. Método experimental específico de séries temporais

Além dos métodos de análise e previsão de séries temporais, esse trabalho utilizou ferramentas descritivas disponibilizadas pela área de estatística.

Com um entendimento do comportamento populacional, pode-se decompor a série para visualizar com mais clareza a sua tendência, sazonalidade, e ruído, se houverem.

3.2.1 Importação e tratamento

A ABECS disponibilizou dados de compra entre janeiro de 2016 até dezembro de 2022, que no total correspondem a 84 registros. Para este espaço de tempo não houve dados ausentes. A Tabela 1 apresenta as variáveis contidas no banco nos dados da ABECS, além das variáveis provenientes do IBGE.

Tabela 1- Descrição do conjunto de dados utilizado no trabalho

origem	índice	variáveis	tipo de dado	Descrição
ABECS	0	Crédito	inteiro	Variação volumétrica de cartão de crédito total no período
ABECS	1	Débito	decimal	Variação volumétrica de cartão de débito total no período
ABECS	2	Pré-Pago	inteiro	Variação volumétrica de cartão de pré-pago total no período
IBGE	3	IPCA	decimal	Variação percentual, ano a ano, do índice de preço ao consumidor amplo
IBGE	4	Desemprego	decimal	Variação percentual, mês a mês, do número de pessoas com idade para trabalhar (acima dos 14 anos) que não estão trabalhando, mas estão disponíveis e tentam encontrar trabalho
IBGE	5	Renda_Media	inteiro	Variação volumétrica, mês a mês, da renda média das pessoas conforme pesquisa nacional por amostra de domicílios contínua (PNAD)
ABECS	6	time	data	Série temporal de mês/ano

Fonte: Autoral (2023).

É válido ressaltar que o espaço de tempo foi definido pela disponibilidade dos dados da ABECS, e não do IBGE; afinal, a ABECS não possui registros superiores a dezembro de 2022. Como a pesquisa envolve a análise da cointegração, a qual segundo A. Nielsen (2021) trata de uma relação real entre duas séries temporais, é necessário que os campos temporais entre ABECS e IBGE sejam coincidentes. Portanto, os dados do IBGE que superiores a dezembro de 2022 não foram utilizados.

3.2.2. Análise exploratória

Com o intuito de compreender melhor os dados analisados, foram construídos alguns gráficos que ilustram o comportamento temporal das variáveis em questão. A Figura 7 apresenta a série histórica do conjunto de dados da ABECS.

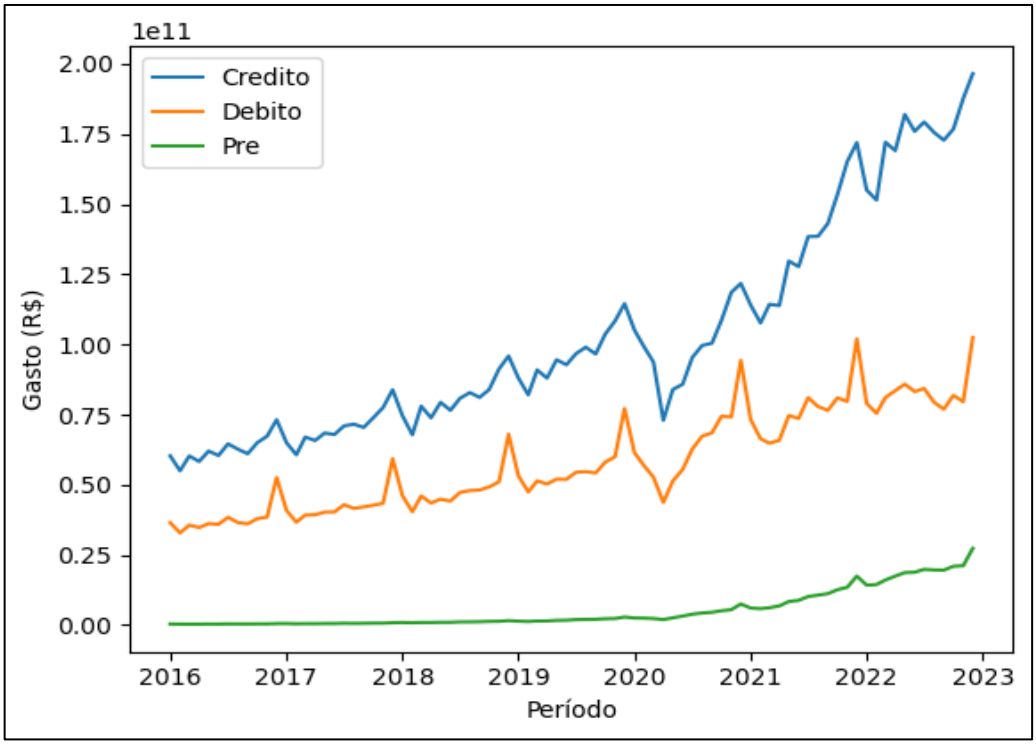

Figura 7 – Série histórica do gasto da população brasileira em Crédito, Débito e Pré-Pago

Fonte: Autoral (2023).

Conforme visto na Figura anterior, há uma notável queda nos gastos a partir de março de 2020, indicando um aparente impacto da pandemia. É muito provável que interferências

como a da pandemia impactem na lógica de predição, portanto, foi necessário realizar um experimento, que será devidamente apresentado mais adiante onde foram substituídos os registros do período pandêmico, por outros obtidos de forma preditiva, desta maneira a tendência comportamental dos anos anteriores pode ser preservada.

Os dados do IBGE também apresentaram um ruído no período pandêmico, assim como nos dados da ABECS. A Figura 8 apresenta a série histórica da variação do IPCA, desemprego e renda mensal.

Figura 8 – Série histórica da renda média mensal, variação percentual do IPCA, e da taxa de desemprego da população brasileira

Fonte: Autoral (2023).

Durante a pandemia, a demanda subiu muito pelos produtos, orientada por um movimento populacional de estocar alimento em função do medo e da incerteza daquele período. Houve muitos estabelecimentos que acabaram fechando, impactados pela redução de

aglomerações de pessoas em virtude das medidas de segurança pública que foram adotadas pela liderança governamental. Por fim, sobre a renda média, é complexo fazer uma análise macro, pois deve-se compreender que mesmo com o aumento da renda média no período pandêmico, isso não necessariamente reflete melhora na saúde financeira da população brasileira. Por exemplo, uma, das várias prováveis causas, para isto ter ocorrido é que somente os negócios/empresas com mais recursos financeiros para suportar o impacto do momento sobreviveram, fazendo com que a média subisse uma vez que as menores rendas sucumbiram ao momento crítico causado pela COVID-19.

Ainda sobre o impacto da pandemia em 2020, a Figura 9 apresenta a sazonalidade sob um ângulo categorizado por mês ao longo dos anos. Com isto é possível observar de forma mais evidente a influência da pandemia.

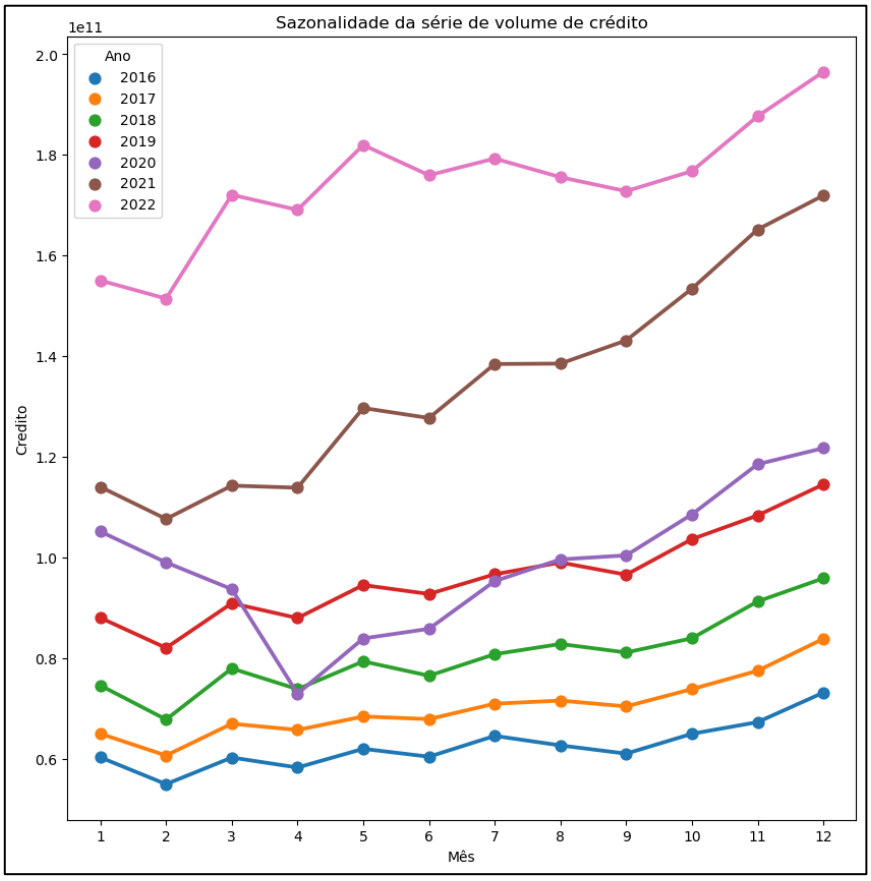

Figura 9 – Série histórica do gasto da população brasileira em Crédito, Débito e Pré-Pago

Fonte: Autoral (2023).

A partir da Figura 9 é possível identificar que após março de 2020, houve uma mudança na sazonalidade. Esta observação aumenta a dificuldade da predição, uma vez que esta mudança

é inesperada; portanto, será necessário tratar os dados influenciados pela pandemia, para que o modelo não seja enviesado pela crise de saúde global.

3.2.3. Análise descritiva estatística

Analisar o comportamento populacional dos dados é uma parte importante e significativa do trabalho, pois é um aliado significativo no entendimento do sistema, permitindo conjecturar hipóteses e criticar os resultados de forma mais contundente.

O gráfico de distribuição de frequência, ou como é mais popularmente conhecido, histograma, ajuda na interpretação do conjunto de dados de uma forma geral. A Figura 10 contém os histogramas das 4 variáveis preditoras utilizadas no trabalho (Crédito, Desemprego, IPCA e Renda Média), exceto pelo tempo, contendo além da distribuição dos dados, sua média (linha vermelha) e mediana (linha amarela).

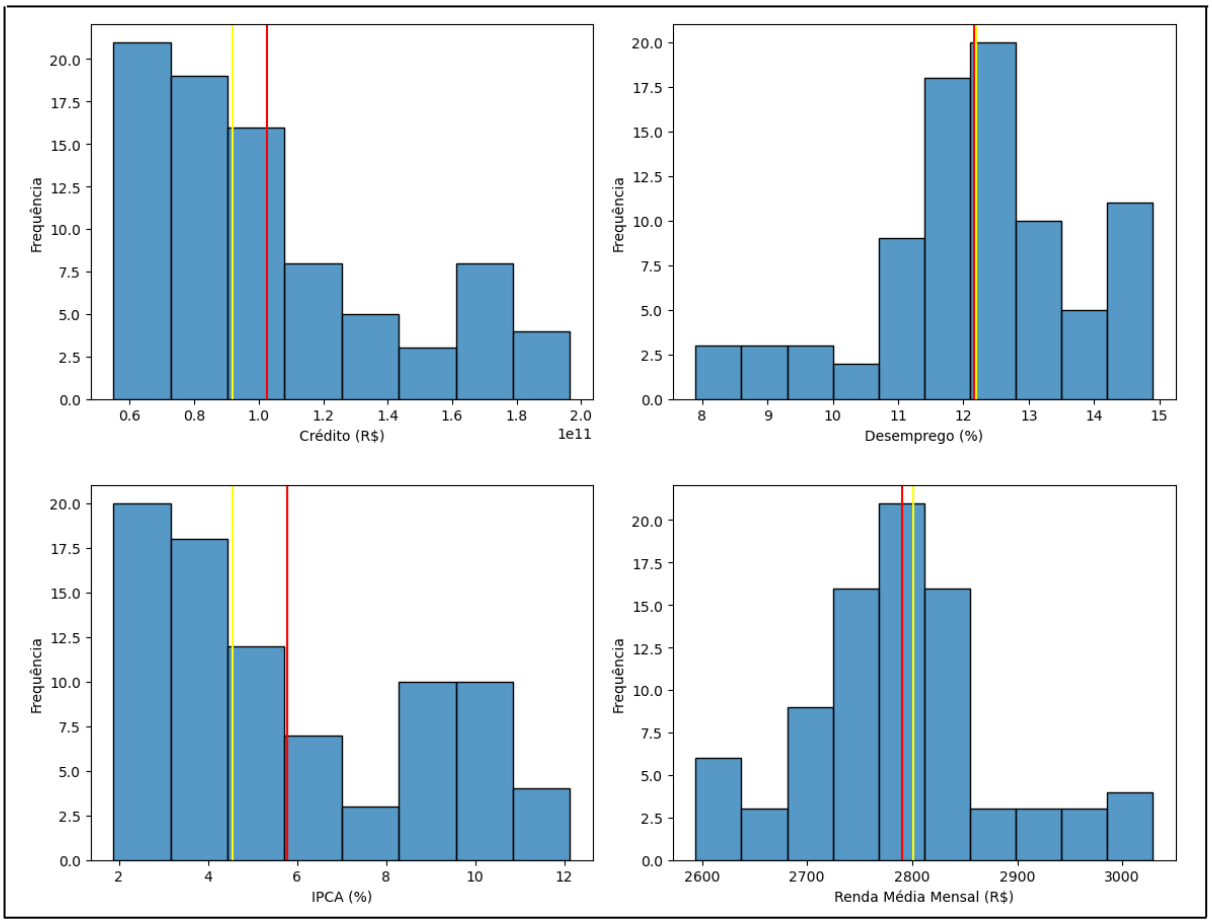

Figura 10 – Histograma do volume de crédito, taxa percentual de desemprego, e IPCA, e renda média mensal da população brasileira

Fonte: Autoral (2023).

Conforme discutido por Pedro A. Morettin e Wilton de O. Bussab (2004), o resumo de dados através de gráficos, como os histogramas da Figura 9, fornecem muito mais informações sobre o comportamento da variável do que a própria tabela original dos dados. A mediana é a medida que ocupa a posição central da série de observações, quando estas estão ordenadas em ordem crescente. Enquanto que a média aritmética é a soma das observações dividida pelo número total delas. O resumo de um conjunto de dados apenas por medidas representativas da posição central da população esconde toda a informação sobre a variabilidade do conjunto de observações. Para observar a variabilidade, um critério frequentemente usado é o desvio padrão, o qual é definido como a raiz quadrada positiva da variância. O desvio padrão indica o erro cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados, no caso a média. A Tabela 2 a seguir apresenta o resumo do estudo posicional dos dados.

Tabela 2 – Resumo de dispersão das variáveis preditoras

medida	Credito	IPCA	Desemprego	Renda _Media
média	102.492.432.112	5,78	12,17	R$ 2.790,00
mediana	91.997.614.158	4,56	12,20	R$ 2.801,00
desvio padrão	38.398.073.987	3,01	1,59	R$ 93,36
mínimo	54.945.522.073	1,88	7,90	R$ 2.594,00
25%	72.589.160.000	3,25	11,40	R$ 2.731,75
50%	91.997.610.000	4,56	12,20	R$ 2.801,00
75%	119.322.610.000	8,73	12,10	R$ 2.837,00
máximo	196.495.616.595	12,13	14,90	R$ 3.029,00

Fonte: Autoral (2023).

Tanto para crédito como para a IPCA, a média é maior do que a mediana, o que indica que a curva gaussiana é assimétrica à direita, enquanto que no desemprego e na renda média, a mediana é superior à média, portanto a curva de Gauss torna-se assimétrica à esquerda. Contudo, a média é muito mais próxima da mediana para os casos da renda média e do desemprego, o que se traduz em uma curva normal mais simétrica comparada ao crédito e ao IPCA. Para Pedro A. Morettin e Wilton de O. Bussab (2004), uma distribuição simétrica, também conhecida como distribuição gaussiana em homenagem ao matemático alemão Carl

Friedrich Gauss (1777 - 1855), é definida quando 50% das observações do conjunto de dados estão distribuídas no intervalo interquartil (1), conforme a Figura 11. Nessa figura pode-se notar uma medida denominada p-quantil, indicada por *q(p)*. Na qual *p* é uma proporção qualquer, 0< *p* <1, tal que 100*(p)*% das observações sejam menores do que *q(p)*. Os quantis mais importantes estão ilustrados na Figura 12.

$$\text{Intervalo interquartil} = q3 - q1 = q(0{,}75) - q(0{,}25) \qquad (1)$$

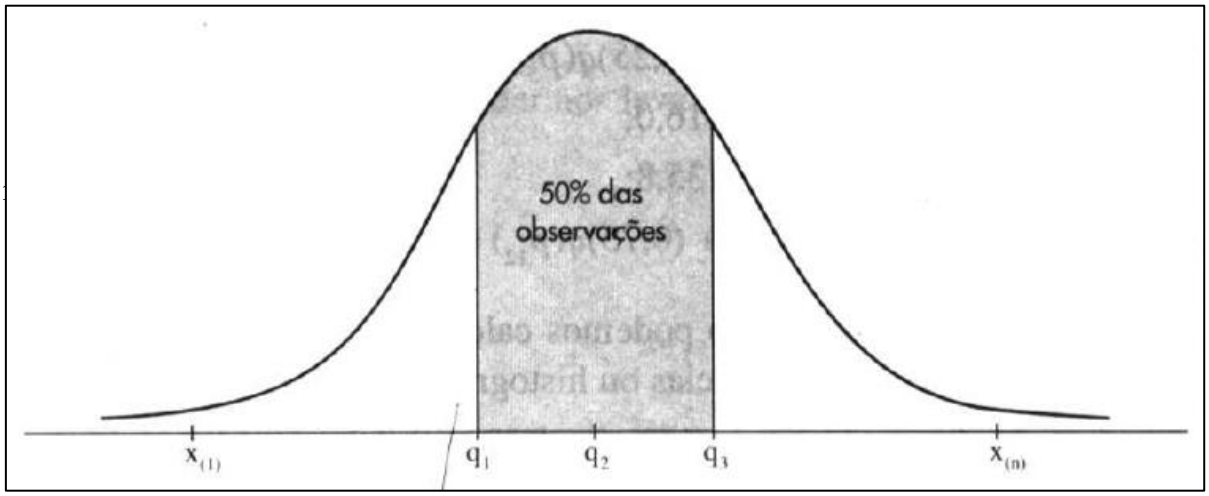

Figura 11 – Distribuição simétrica: normal ou gaussiana

Fonte: Pedro A. Morettin e Wilton de O. Bussab (2004)

$q(0{,}25)$: 1º Quartil = 25º Percentil
$q(0{,}50)$: Mediana = 5º Decil = 50º Percentil
$q(0{,}75)$: 3º Quartil = 75º Percentil
$q(0{,}40)$: 4º Decil
$q(0{,}95)$: 95º Percentil

Figura 12 – Principais quantis utilizados

Fonte: Pedro A. Morettin e Wilton de O. Bussab (2004)

A importância desta análise deve-se ao fato de que quanto mais simétrica for a curva de Gauss, mais próximo de assumir a hipótese de que 99,3% da distribuição dos dados estará contida entre o limite superior (LS) e o limite inferior (LI) da distribuição dos dados, conforme ilustrado pela Figura 13.

Embora o crédito não seja tão simétrico quando comparado às outras variáveis, não significa que o estudo será afetado, a curva de crédito ilustrada pela Figura 7 mostra que a medida que o tempo passa, a demanda por crédito tem uma tendência de crescimento, ou seja, é de se esperar que com o passar do tempo, a frequência dos valores menores de crédito fiquem cada vez mais distantes da mediana e isso naturalmente se traduz na assimetria positiva, que é quando a curva se alonga à direita.

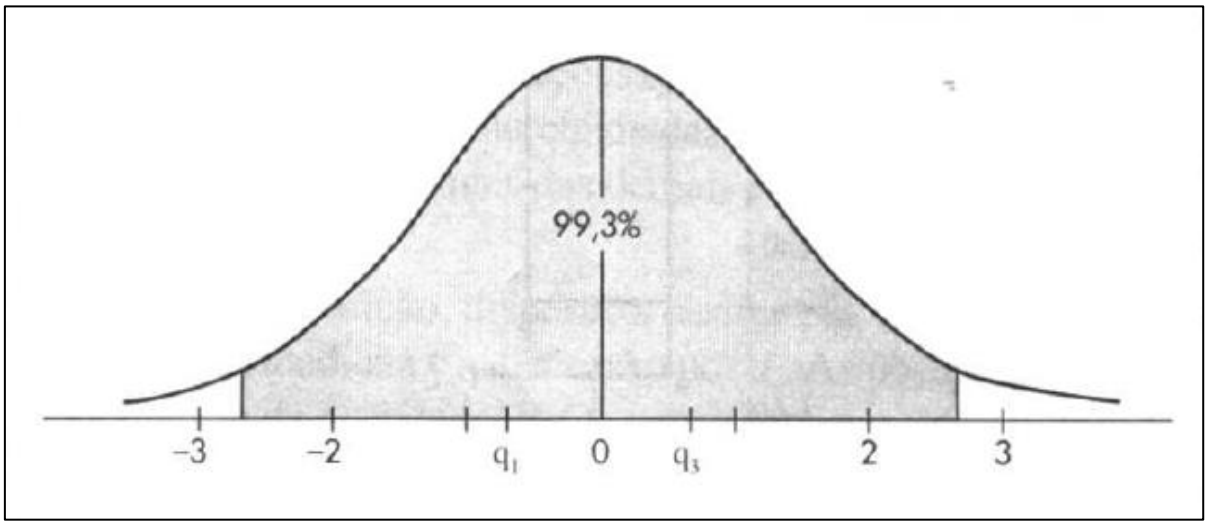

Figura 13 – Área sob a curva normal entre LI e LS

Fonte: Pedro A. Morettin e Wilton de O. Bussab (2004)

Após identificar o comportamento crescente da série temporal, o mesmo fenômeno deve ser verificado com mais profundidade na decomposição da série temporal, buscando compreender as suas principais características conforme o capítulo 2.4..

3.2.4. Decomposição da série temporal de volume de crédito

A decomposição é uma importante etapa na modelagem de séries temporais, permitindo constatar tendência, sazonalidade e também a aleatoriedade do sistema. A Figura 14 ilustra a decomposição da série temporal de crédito. Na qual pode-se verificar que há, de fato, uma tendência de crescimento positiva. Também é possível notar que o ângulo da curva aumenta ao final de 2020, indicando um aumento na demanda por volume de crédito no sistema ao longo do tempo.

A sazonalidade apresentada possui um formato bem definido, com picos claros de volume de crédito ao longo de 12 meses. Contudo, o impacto da pandemia fica evidente ao se

observar a queda aguda do resíduo no começo de 2020, essa característica, conforme visto na Figura 4, corresponde a um fenômeno externo ocorrido que gerou impacto no sistema observado. Casos como esse não são bons para o modelo preditivo, uma vez que o fator da COVID-19 não possui qualquer relação cronológica para ocorrer novamente, logo o histórico fica prejudicado. Tentar predizer os meses futuros com este fator incluso no histórico seria desvantajoso para o trabalho.

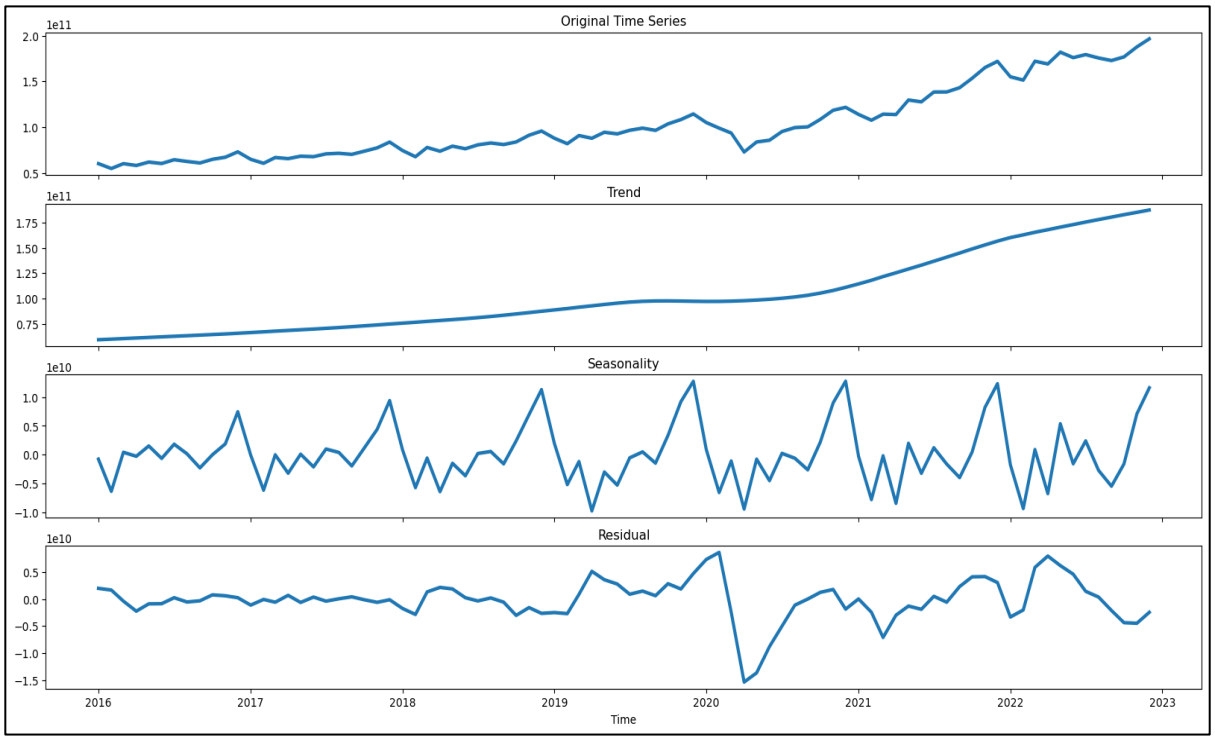

Figura 14 – Decomposição da série temporal de Crédito

Fonte: Autoral (2023)

3.2.6. Teste de anomalia e correção temporal

Uma vez identificado o impacto de um fator externo no sistema, é necessário compreender quando, especificamente, este fator começou a influenciar e até que ponto seguiu impactando. Para isso KATS disponibiliza uma ferramenta de detecção de anomalia temporal, chamada *RobustStatDetector*. Este algoritmo detecta a alteração de pontos na série temporal através de deslocamentos médios. A série temporal é suavizada através de uma média móvel, calculando as diferenças da série temporal para um número fixo de pontos, podendo ser configurado caso necessário. Em seguida, os *z-scores* e o *p-value* são calculados para cada ponto da série. Quando o *p-value* do ponto em questão for inferior ao *p-value-cutoff* calculado

pelo algoritmo, há uma anomalia temporal detectada. A Figura 15 demonstra as 2 anomalias detectadas pelo algoritmo no sistema de volume de crédito.

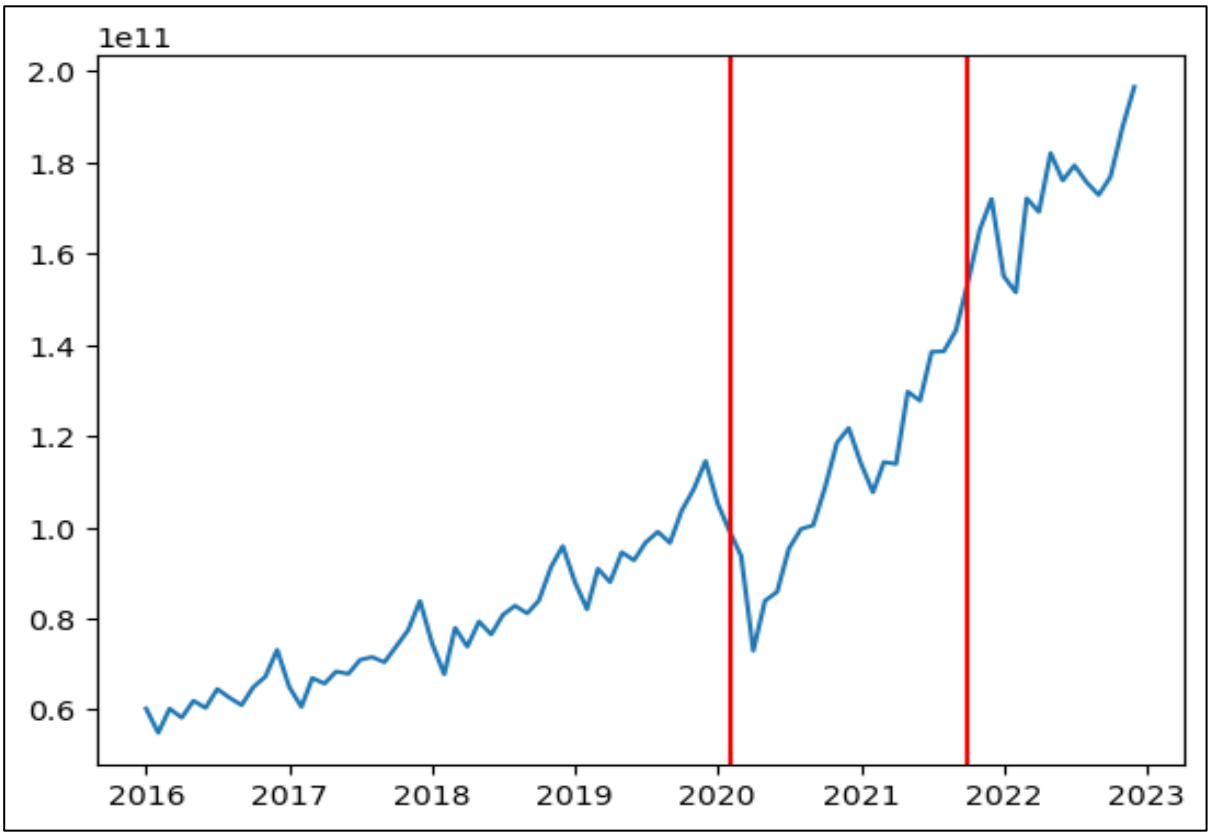

Figura 15 – Anomalias temporais na série temporal de volume de crédito

Fonte: Autoral (2023)

A primeira anomalia, foi detectada em fevereiro de 2020, o que é plausível tendo em vista que neste período os veículos de informação já alertavam para os casos de COVID-19 em partes do mundo, principalmente na China. Em 11 de março de 2020 a OMS (Organização Mundial da Saúde) declarou o estado de pandemia.

Já a segunda anomalia temporal foi detectada em outubro de 2021. Segundo o site da Agência Brasil, disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2021-12/ (Acessado em: 01 de Maio de 2023), no período da anomalia temporal em questão, foi registrado uma queda de aproximadamente 90% no número de óbitos por COVID-19, com o advento das vacinas em março de 2021. O número de mortes começou a regredir com o passar dos meses e as pessoas aos poucos voltaram a sair, comércios reabriram, e a rotina foi cautelosamente sendo reestabelecida, portanto, a detecção no aumento indicada pela segunda anomalia também é válida perante os fatos ocorridos naquele período.

Para ilustrar mais claramente o impacto do período pandêmico na série temporal que está sendo tratada, a Figura 16 demonstra o comportamento da curva de mortes por COVID-19 em ordem cronológica até fevereiro de 2023. É válido ressaltar que apenas a média móvel de mortes não é suficiente para realizar qualquer afirmação quanto ao comportamento da curva volume de crédito para o período, pois como o próprio presente trabalho busca identificar, existem outros fatores a serem considerados que podem influenciar na variação nos gastos da população. Contudo, ela ajuda a compreender com mais confiabilidade as anomalias detectadas.

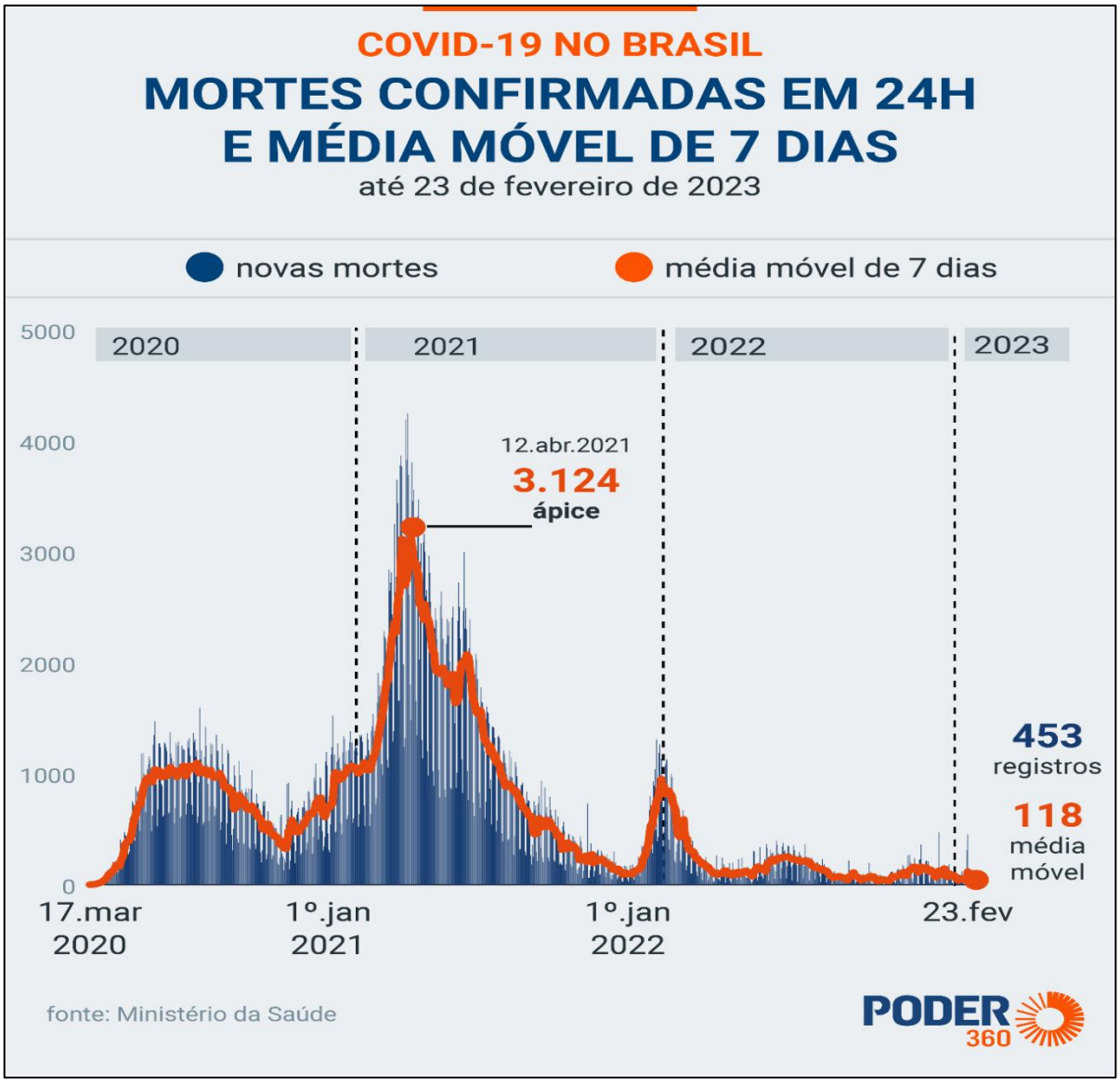

Figura 16 – Média móvel de vítimas de pandemia

Fonte: Poder Data 360 em parceria com o Ministério da Saúde (2023)

Como explorado anteriormente, o impacto gerado pela pandemia possui grande chance de influenciar negativamente para o modelo. Para evitar que este fator prejudique o trabalho, a

série temporal foi cortada de fevereiro de 2020 até fevereiro de 2022. Os dados ausentes neste período foram substituídos pelos dados previstos pelo próprio aprendizado de máquina do trabalho. Posteriormente, com os valores previstos pelo modelo, utilizando a base entre janeiro de 2016 até janeiro de 2022 foi ajustado um modelo suprimindo a influência da pandemia. Com a previsão deste modelo, formou-se uma nova base de dados com as observações a partir de março de 2022 reintegrados a base. Em seguida, foi construído um novo modelo de predição para conclusão do projeto. A série temporal com o corte descrito anteriormente está ilustrada na Figura 17. permitindo notar uma tendência linear e sazonalidades mais bem definidas.

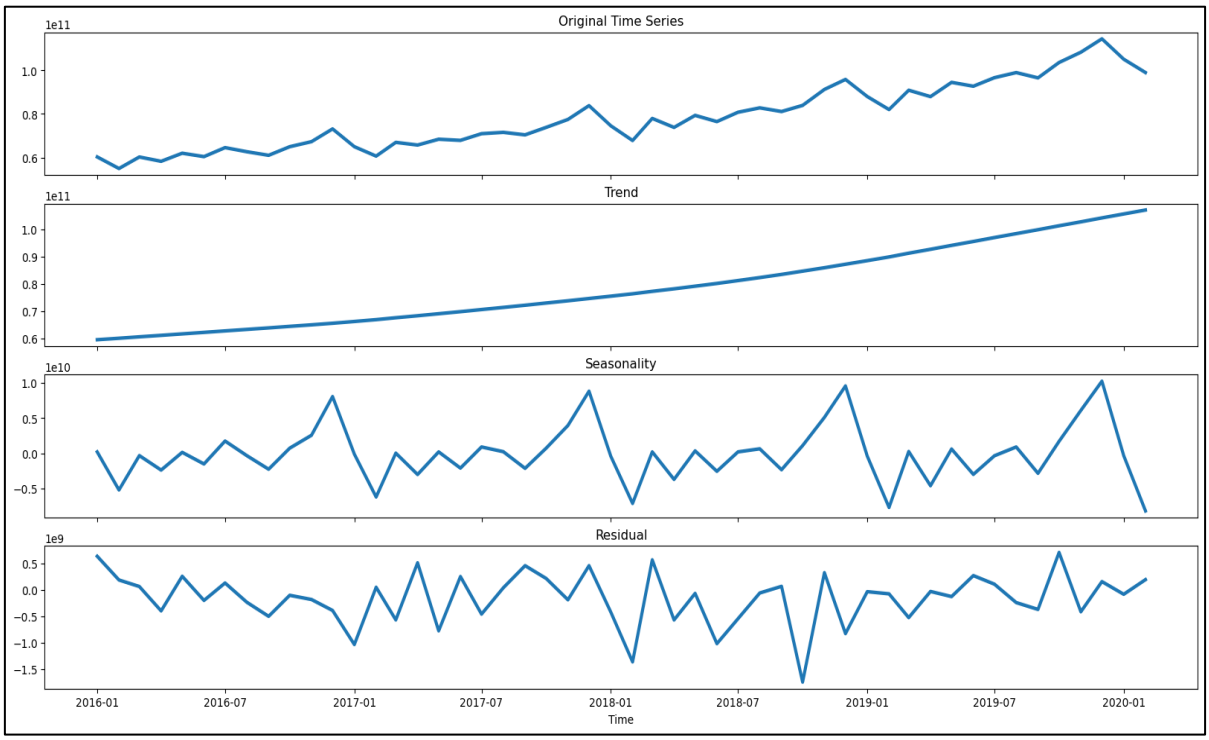

Figura 17 – Anomalias temporais na série temporal de volume de crédito

Fonte: Autoral (2023)

Por fim, antes de seguir para a construção do modelo em si, ainda é necessário compreender mais alguns fatores sobre a série temporal, como por exemplo sua estacionariedade.

3.2.7. Compreendendo a estacionariedade

Muitos modelos estatísticos comumente utilizados dependem de uma série temporal estacionária. Sob uma perspectiva prática, uma série temporal estacionária é aquela onde as

propriedades estatísticas são razoavelmente estáveis ao longo do tempo, sobretudo no que diz respeito a média e à variância.

Dentre os principais testes utilizados para avaliar estacionariedade, pode-se citar Dickey-Fuller Aumentado (ADF) o qual postula a hipótese de que uma raiz unitária está presente na série. Dependendo dos resultados do teste, essa hipótese nula (H0) pode ser rejeitada para um nível de significância especificado, o que significa que a presença de um teste de raiz unitária pode ser rejeitada em um determinado grau de confiabilidade, indicando que a série é estacionária, ou o teste pode aceitar H0, o que reflete em uma série não estacionária.

Para o sistema estudado, de acordo com o teste estatístico da ADF, é possível dizer com 95% de confiabilidade de que o sistema não é estacionário, pois a estatística ADF é maior do que os valores críticos, o que acaba não rejeitando H0. Os resultados estão dispostos na Tabela 3. Esse resultado era esperado porque a partir da Figura 7, era possível notar o incremento da média e também da variância com base na curva da série temporal, indicando a existência de uma tendência de evolução.

O problema é que um modelo de série temporal não estacionário sofrerá variações em relação a sua média. Ou seja, o viés e o erro do modelo variarão ao longo do tempo, chegando ao ponto que o resultado obtido pelo aprendizado de máquina se torna questionável.

É comum que uma série temporal possa se tornar estacionária o suficiente com algumas transformações simples. Uma transformação *log* e uma transformação raiz quadrada são bastante utilizadas nesse contexto. Da mesma forma, a remoção de uma tendência é normalmente feita por diferenciação, que consiste em subtrair o valor atual pelo valor prévio. Há situações em que a série deve ser diferenciada mais de uma vez; porém, é importante destacar que, se diferenciação for muito utilizada (mais de duas ou três vezes), é improvável que se consiga solucionar o problema da estacionariedade com a diferenciação.

Tabela 3 – Resultados para a série temporal original

métrica	resultado
ADF Statistic	1,708
p-value	1,000
Lags	11,000
critical value 1%	-4,219
critical value 5%	-3,533
critical value 10%	-3,198
H0	ACEITA

Fonte: Autoral (2023)

Quanto mais negativo for o valor da estatística ADF, mais estacionária é a série. Nesse sentido, realizou-se 3 experimentos para tornar a série estacionária para seguimento do estudo, onde os resultados estão evidenciados na Tabela 4.

Tabela 4 – Resultados para a série temporal com diferenciação [-1]

métrica	Resultado diff[-1]	Resultado log diff[-1]	Resultado raiz ² diff[-1]
Estatística ADF	-18,867	-19,714	-19,786
p-value	0,000	0,000	0,000
Lags	10,000	10,000	10,000
critical value 1%	-4,219	-4,219	-4,219
critical value 5%	-3,533	-3,533	-3,533
critical value 10%	-3,198	-3,198	-3,198
H0	REJEITADA	REJEITADA	REJEITADA

Fonte: Autoral (2023)

Conforme os resultados expostos anteriormente, apenas uma única aplicação de diferenciação foi suficiente para transformar a série de não-estacionária em estacionária. Todos os métodos empregados mostraram-se eficazes, porém o método adotado para deixar a série

estacionária foi o logarítmico por facilitar a visualização gráfica. Contudo, é válido ressaltar para fins didáticos que melhores resultados são obtidos quanto mais negativo for o valor da estatística ADF. A Figura 18 ilustra como é o formato de uma série temporal após passar pela diferenciação e transformar-se em série uma estacionária.

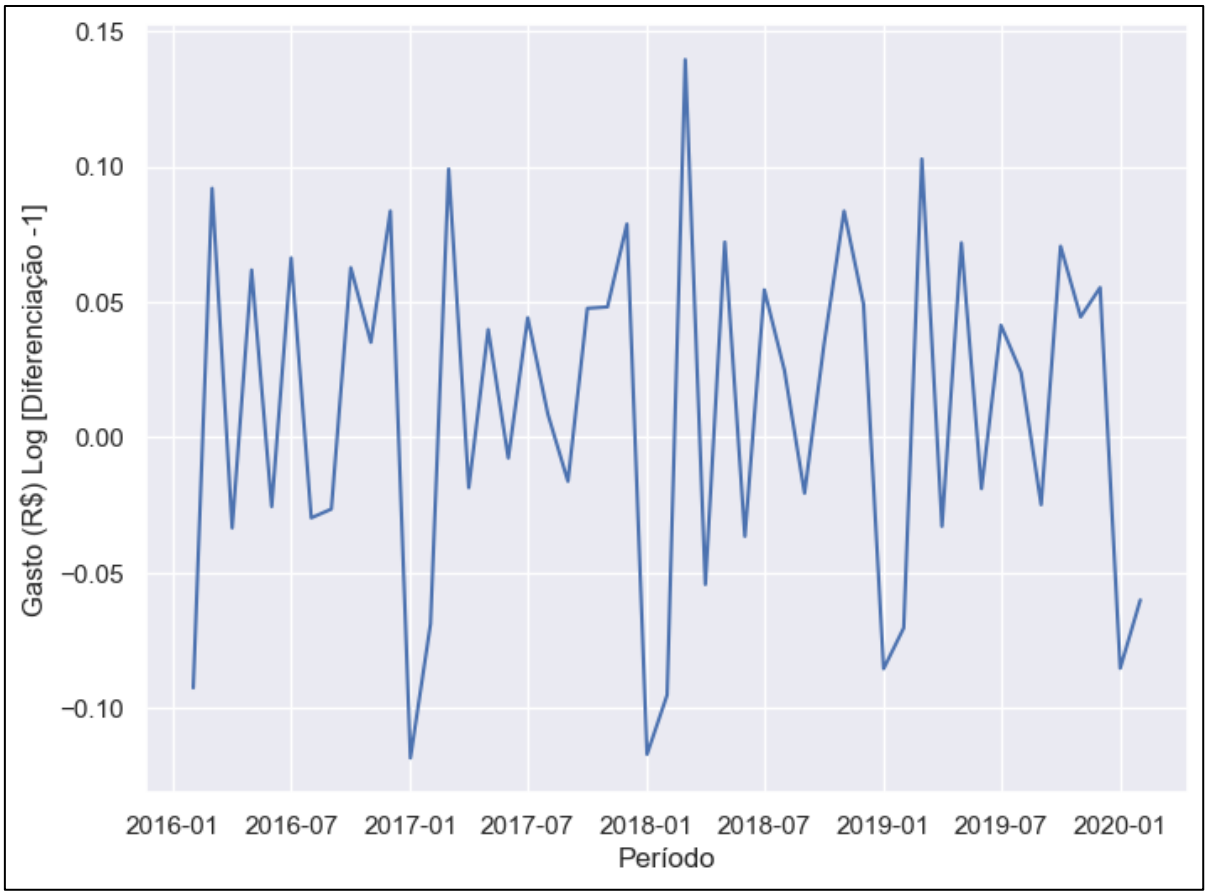

Figura 18 – Série temporal de volume de crédito transformada em estacionária através do método logarítmico com diferenciação 1

Fonte: Autoral (2023)

Com a estacionariedade implementada, agora o próximo passo é entender o conceito da autocorrealção da série, que permite informar se existem pontos temporais do sistema que se relacionam com valores em outros pontos do sistema.

3.2.8. Compreendendo a autocorrelação

A autocorrelação, em resumo, busca verificar se o valor de uma série temporal em um determinado ponto no tempo pode estar correlacionado com o valor de outro ponto no tempo.

Para clarificar o conceito, considere que numa série temporal anual, cujos dados diários de temperatura são registrados. Percebe-se ao analisar os dados que os dias 20 de maio mais quentes de cada ano tendem a se correlacionar com os dias 20 de agosto mais frios. Portanto, há possibilidade de utilizar este conhecimento adquirido para predizer a temperatura do dia 20 de agosto.

Conforme Aileen Nielsen (2021), a questão principal deste tema é – "Existe uma correlação entre quaisquer dois pontos em uma série temporal específica com uma distância fixa em particular entre eles?". Em outras palavras, o objetivo desta análise é identificar a correlação linear de um sinal no tempo com uma cópia atrasada sobre si mesmo. Informalmente, seria a semelhança entre observações em função do intervalo de tempo entre elas.

O estudo da autocorrelação é especialmente importante para modelos autorregressivos, pois são determinantes para a definição de parâmetros do modelo, que serão discutidos mais adiante.

Existem duas funções que ajudam a construir o gráfico de autocorrelação, são elas: *autocorrelation function* (ACF) e *partial autocorrelation function* (PACF). As regiões críticas tanto para a PACF como para a ACF são iguais, pois seus limites são determinados por $\pm 1.96\sqrt{1/n}$, onde n é o número de amostras. Esta é uma regra estatística para determinar uma estimativa diferente de 0 que seja significativa, refletindo a região crítica mencionada anteriormente. Essa regra se baseia em um tamanho de amostra grande o bastante e em uma variância finita para o processo. A Figura 19 apresenta os gráficos ACF e PACF para a série temporal de estudo (Alencar, A. P. Rocha, 2009).

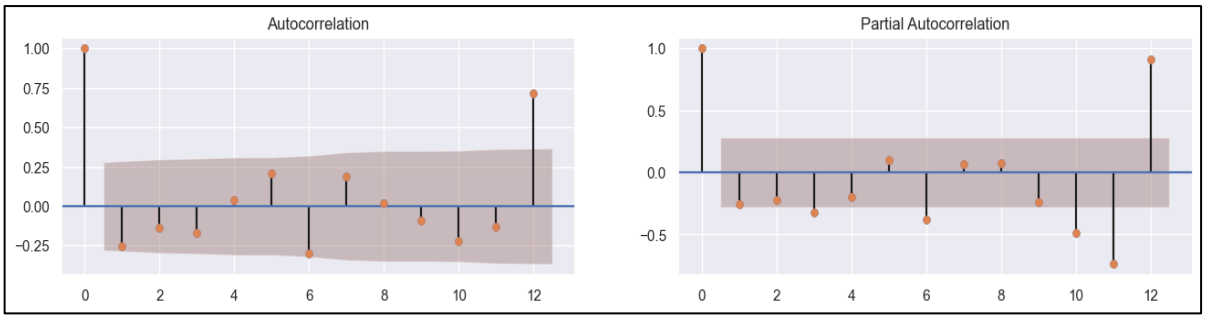

Figura 19 – Gráfico ACF e PACF para 12 lags da série temporal de volume de crédito

Fonte: Autoral (2023)

Ambas as funções iniciam com lag=0, que apresenta correlação máxima, igual a 1. Isso sempre acontece pois trata-se da correlação da série com ela mesma. A diferença entre a ACF e a PACF é a inclusão ou exclusão de correlações indiretas no cálculo. A PACF de uma série

temporal para um determinado lag é a correlação parcial dessa série temporal com ela mesma nesse lag, dadas todas as informações entre os dois pontos.

É possível notar uma área sombreada no gráfico, trata-se do intervalo de 95% de confiabilidade das funções. Tudo que está imerso nessa região crítica, significa que estatisticamente que está próximo de 0, e tudo que está fora dela é estatisticamente diferente de 0. As seguintes observações podem ser feitas ao analisar a Figura 20:

1. Existem muitos pontos dentro do intervalo de confiança, tanto do ACF como do PACF, portanto a série temporal não é aleatória;
2. Tanto na ACF como na PACF há uma correlação forte no lag=12. A partir da Figura 20, pode-se identificar a ordem da série temporal através de um modelo autoregressivo (AR[p=12]).

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off (Geometric decay)	Significant at lag q / Cuts off after lag q	Tails off (Geometric decay)
PACF	Significant at each lag p / Cuts off after lag p	Tails off (Geometric decay)	Tails off (Geometric decay)

Figura 20 – Referência para enquadramento de tipo de série temporal

Fonte: L. Monigatti (2022)

Para concluir, com os parâmetros para modelos autorregressivos definidos, já pode-se realizar a construção dos modelos e avaliar seus resultados para uma primeira predição, que tem por objetivo corrigir o viés gerado pela pandemia. Após obter novos dados que não estejam enviesados pela COVID-19, calcula-se o modelo final, incluindo as investigações com as variáveis econométricas.

4. RESULTADOS E DISCUSSÕES

Os resultados apresentados nesta seção buscaram analisar a base de dados apresentada anteriormente visando extrair informações e padrões implícitos que podem auxiliar na compreensão da influência do IPCA e da taxa de desemprego no gasto de crédito da população brasileira. Com no referencial teórico e no conjunto de dados devidamente tratado, foi possível elaborar alguns modelos e comparar seus resultados para dar seguimento ao estudo.

É importante destacar que os dados analisados incluem o período da pandemia de COVID-19. Por esta razão, foi planejado um cenário experimental com a aplicação de uma abordagem para estimar valores que reduzissem o impacto da pandemia na série histórica. Esse cenário foi necessário para suavizar os dados para que a previsão não fosse enviesada devido às mudanças bruscas de comportamento durante a pandemia.

4.1. **Cenário experimental para redução do impacto da COVID-19**

Existem diversos modelos que abrangem o campo das séries temporais que podem ser testados; contudo, para esta pesquisa o foco se deu em 3 algoritmos devido à alta aplicabilidade para diversos conjuntos de dados.

Para a classe autoregressiva experimentou-se o SARIMA e o Holt-Winters (HW), enquanto que para a classe aditiva foi utilizado o Prophet. Para analisar e comparar a eficácia dos modelos, foi utilizado como base de treino o conjunto de dados entre janeiro de 2016 até janeiro de 2019, e para testar utilizou-se os dados entre fevereiro de 2019 até fevereiro de 2020, totalizando 1 ano. Os resultados dos modelos treinados podem ser vistos na Figura 21, assim como suas predições para o ano seguinte.

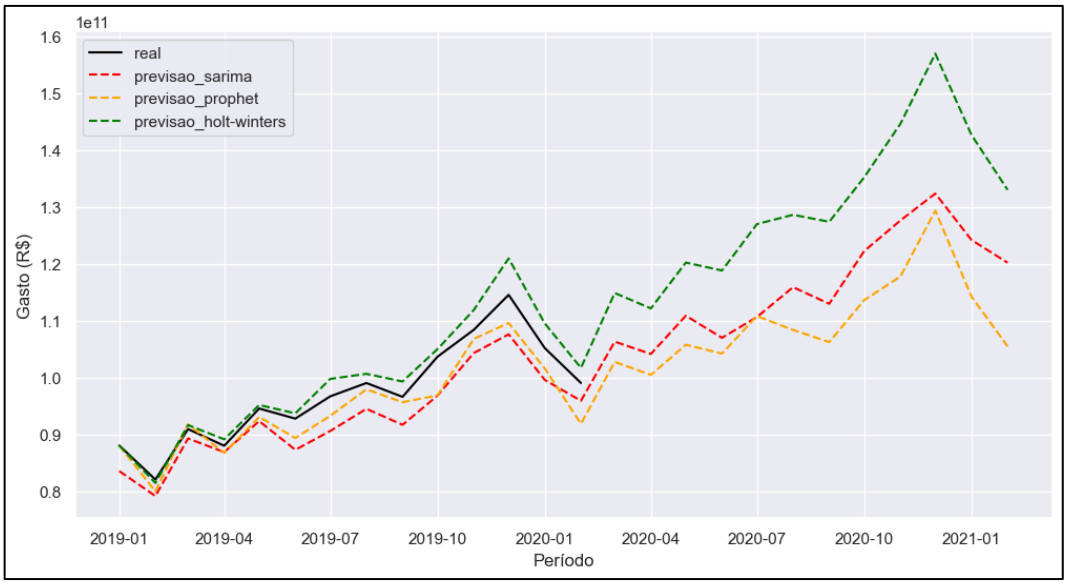

Figura 21 – Resultado das predições para o volume de crédito

Fonte: Autoral (2023)

O gráfico da Figura 21 sugere que tanto Prophet quanto SARIMA apresentam resultados abaixo dos valores reais, enquanto que HW está, na maior parte, acima. É importante destacar que, com base nesses resultados apenas, não é possível afirmar qual modelo apresenta os melhores resultados. Considerando que todos os modelos mantêm um comportamento visual semelhante aos dados analisados, pode-se escolher aquele que melhor se aproxima do objetivo da análise.

Para contribuir com a decisão de escolha do modelo, além da análise visual, utilizou-se o Erro Percentual Absoluto Médio (Mean Absolute Percentage Error - MAPE), que representa o Desvio Absoluto Médio (Mean Absolute Deviation - MAD) em escala percentual. Esse tipo de métrica é importante para inibir que os sinais (positivos e negativos) impactem no cálculo do desempenho, mantendo o erro absoluto e gerando um erro relativo absoluto que permite uma fácil comparação entre séries temporais. A Tabela 5 ilustra esses resultados.

Tabela 5 – Resultados do MAPE para os modelos experimentados

modelo	MAPE
Holt-Winters	2,07%
Prophet	2,77%

SARIMA	4,35%

Fonte: Autoral (2023)

Com a ajuda do MAPE é possível observar que o modelo SARIMA errou mais na base de treino do que os modelos Prophet e HW. Contudo, mesmo analisando essa métrica, ainda fica difícil decidir com base em erros relativos tão próximos entre os modelos. Para contribuir mais com esse processo de decisão, os novos dados foram incluídos conforme ilustrados na Figura 22. De maneira resumida, buscou-se observar uma adequação intermediária entre os dados anteriores a março de 2020 e posteriores a fevereiro de 2021.

Esses dados auxiliam no processo de escolha do modelo, permitindo visualizar suas contribuições no ajuste da série temporal completa. As curvas proporcionadas por SARIMA e Prophet se destacam por assumir uma posição intermediária entre os dados 'real inicial' e 'real final', i.e., esses modelos preenchem a série com menos interferência do que a predição do HW. Levando em consideração que o MAPE do Prophet foi melhor que o do SARIMA, então decidiu-se utilizar o Prophet para substituição dos dados do período pandêmico.

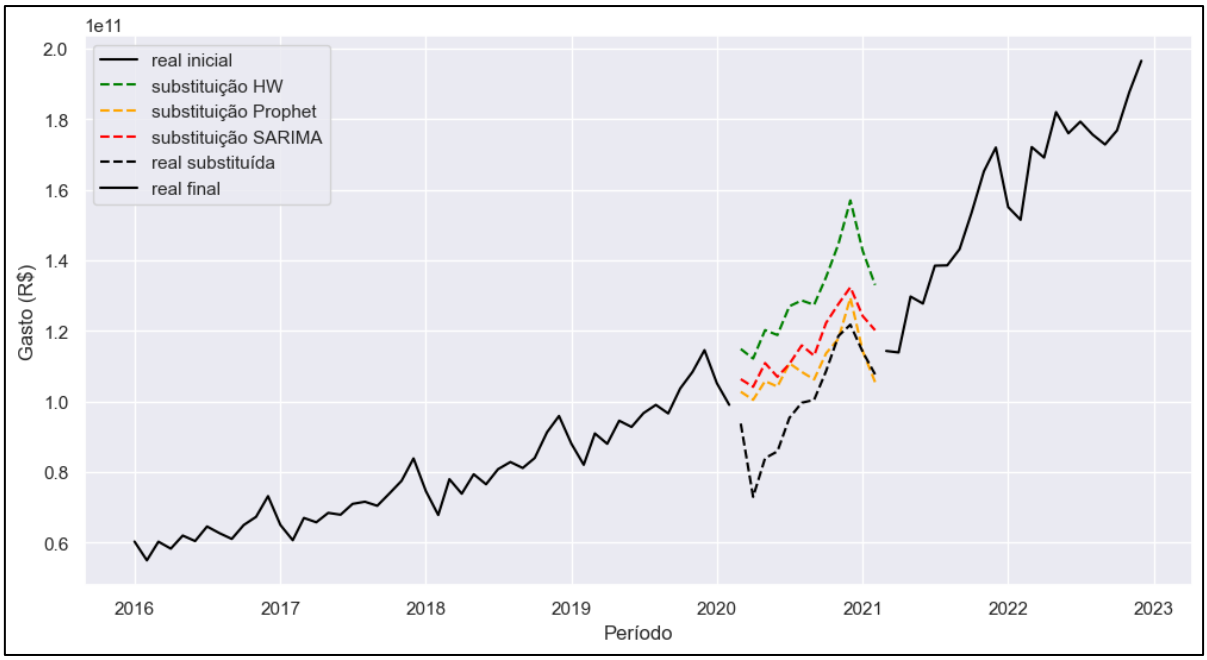

Figura 22 – Possíveis cenários para substituição de viés pandêmico

Fonte: Autoral (2023)

O novo cenário do conjunto de dados com a predição do modelo Prophet entre março de 2020 até fevereiro de 2021, está definido. Embora ainda exista uma certa instabilidade na

sazonalidade após fevereiro de 2021, pode-se formular um modelo inicial e avaliar a sua covariação com as métricas IPCA e taxa de desemprego de forma mais concisa do que antes.

4.2. Aplicação de modelos univariados

A partir da transformação realizada na série temporal, pode-se efetuar uma nova construção de modelos univariados para prever o volume de crédito para os anos subsequentes. Para isso, utilizou-se o mesmo conjunto de algoritmos utilizados anteriormente. A métrica de desempenho do modelo seguiu sendo o MAPE e seus resultados são apresentados pela Tabela 6.

Tabela 6 - Resultados dos modelos experimentados em base normalizada

modelo	MAPE
Holt-Winters	6,86%
Prophet	10,15%
Sarima	6,54%

Fonte: Autoral (2023)

Os resultados sugerem Holt-Winters e Sarima como melhores modelos preditivos. Porém, para uma melhor interpretação, foram ilustradas na Figura 23 as curvas de cada modelo. Além das curvas, também foram adicionados 3 novos pontos de dados provenientes da ABEC's, devido a atualização e disponibilização trimestral feita pela organização. Estes 3 novos períodos de dados referem-se a janeiro, fevereiro e março de 2023 e contribuem para uma decisão mais apurada sobre qual modelo usar.

Figura 23- Cenários previstos pelos modelos em base normalizada

Conforme exemplificado acima, considerando os modelos univariados, o melhor desempenho foi obtido com Sarima devido ao seu comportamento nos dados de teste. Nota-se que sua previsão ficou próxima, mas abaixo do real, sendo mais aconselhável para cenários de negócios. isto é vantajoso pois é mais aconselhado prever menos e entregar mais, do que prever mais e entregar menos. Em relação aos dados novos inseridos, é possível observar que o Sarima seguiu portando-se bem, com um ajuste bem próximo da realidade, concedendo mais confiança para utilizá-lo.

4.3. Aplicação de modelos multivariados

Uma vez que foi possível criar um modelo útil para prever as demandas de crédito, pode-se utilizá-lo como referência para comparar com modelos multivariáveis e, por fim, determinar se os demais campos econométricos contribuem para uma melhor previsão do volume de crédito.

Para esta etapa, serão utilizados 2 grupos de algoritmos multivariados no próprio ambiente da KATS com a possibilidade de análise multivariada: Prophet (com regressores extras) e VAR (Vetores Auto Regressivos). A estrutura lógica dos modelos é diferente, mas sua essência é a mesma. Em outras palavras, ambos buscam aprender como uma série de dados

pode impactar na outra, seguindo o conceito da covariância explicado na revisão bibliográfica. O VAR é uma extensão do modelo auto-regressivo (AR), que visa capturar interdependências lineares entre as múltiplas variáveis adotadas. Em paralelo, o Prophet segue com seu mecanismo aditivo de previsão explicado anteriormente. Contudo, desta vez foram adicionados regressores extras que impactam no modelo preditivo de interesse, permitindo estudar a alteração de comportamento novas séries de dados são adicionadas. Os resultados MAPE de cada experimento estão apresentados na Tabela 7.

Tabela 7- Resultados dos modelos multivariáveis

modelo	qnt. de variáveis	MAPE
VAR - CRÉDITO NORM.; IPCA	2	9,19%
VAR - CRÉDITO NORM.; Renda Média	2	7,94%
VAR - CRÉDITO NORM.; Desemprego	2	8,08%
VAR - CRÉDITO NORM.; IPCA; Renda Média	3	5,97%
VAR - CRÉDITO NORM.; IPCA; Desemprego	3	9,56%
VAR - CRÉDITO NORM.; Renda Média; Desemprego	3	10,63%
VAR - CRÉDITO NORM.; IPCA; Renda Média; Desemprego	4	8,04%
PROPHET - CRÉDITO NORM.; IPCA	2	11,14%
PROPHET - CRÉDITO NORM.; Renda Média	2	14,73%
PROPHET - CRÉDITO NORM.; Desemprego	2	14,73%
PROPHET - CRÉDITO NORM.; IPCA; Renda Média	3	14,73%
PROPHET - CRÉDITO NORM.; IPCA; Desemprego	3	14,73%
PROPHET - CRÉDITO NORM.; Renda Média; Desemprego	3	14,73%
PROPHET - CRÉDITO NORM.; IPCA; Renda Média; Desemprego	4	11,14%

Fonte: Autoral (2023)

Com base nos resultados dos MAPEs, é possível dizer que o algoritmo VAR teve uma performance superior ao Prophet. Mesmo com o Prophet atribuindo regressores extras, o modelo realizou uma previsão idêntica para os seguintes casos:

a. Crédito X Renda Média;
b. Crédito X Desemprego;
c. Crédito X IPCA X Renda Média;
d. Crédito X IPCA X Desemprego;
e. Crédito X Renda Média X Desemprego.

O resultado exposto pelo Prophet é negativo à hipótese de que as variáveis econométricas permitem obter um resultado preditivo melhor do que o modelo univaridado para prever o crédito. A Figura 24 ilustra as curvas distintas obtidas pelos modelos comparadas à realidade.

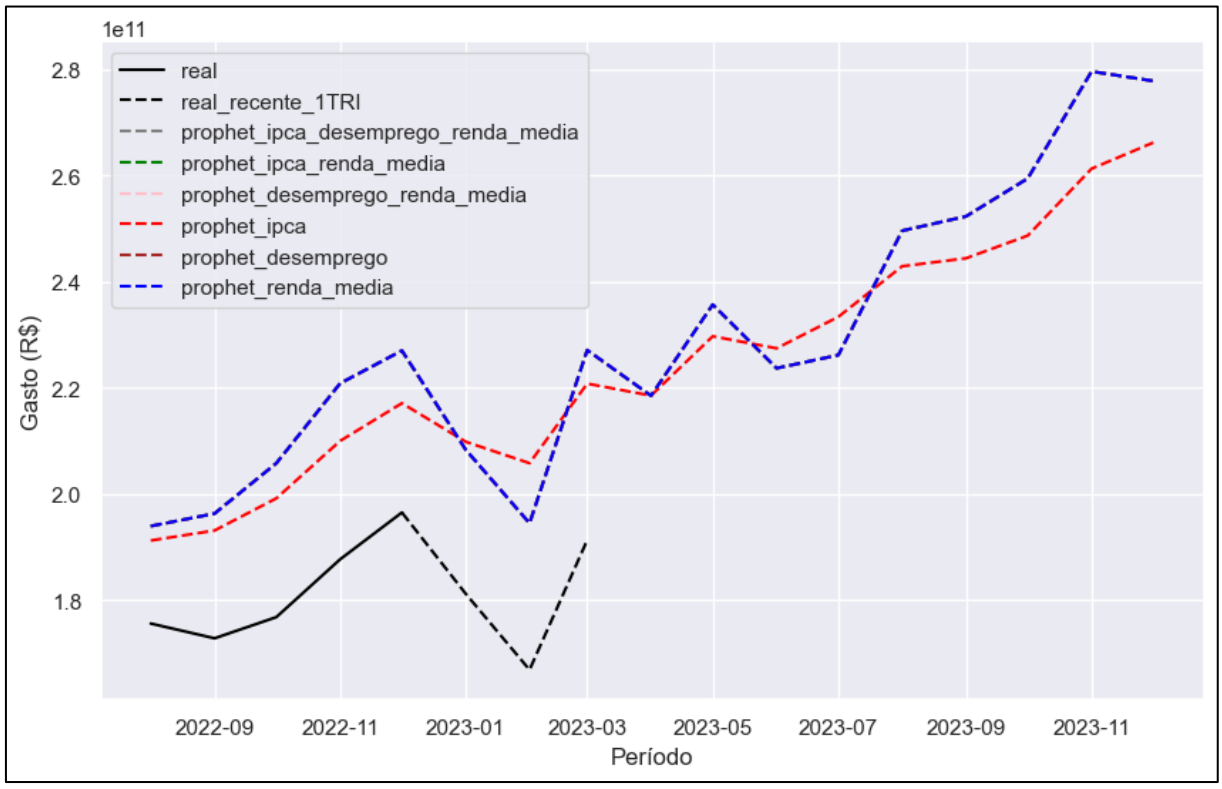

Figura 24 – Cenários previstos pelos modelos Prophet multivariáveis em base normalizada

VAR conseguiu captar melhor a covariância entre as variáveis, gerando cenários mais distintos entre si. Ao analisar a quantidade de variáveis, tem-se o seguinte resumo exposto na Tabela 8.

Tabela 8 - Melhores modelos multivariáveis agrupados por quantidade de variáveis

melhores modelos multivariáveis	qnt. de variáveis	MAPE
VAR - CRÉDITO NORM.; Renda Média	2	7,94%
VAR - CRÉDITO NORM.; IPCA; Renda Média	3	5,97%
VAR - CRÉDITO NORM.; IPCA; Renda Média; Desemprego	4	8,04%

Fonte: Autoral (2023)

Atribuindo somente uma regressão, além dos dados de crédito, o melhor resultado obtido foi com a adição da variável Renda Média. Ao atribuir 2 regressões, além da de crédito, o melhor resultado obtido foi do conjunto IPCA com Renda Média, reduzindo o erro em 1,97%. Por fim, após atribuir as 3 variáveis econométricas para trabalhar em conjunto, cenário único, o resultado do MAPE é o pior entre os citados. A Figura 25 ilustra as curvas geradas por cada um dos modelos citados previamente.

Visando compreender o motivo de algumas variáveis não se saírem tão bem com outras, gerou-se a matriz de correlação de Pearson, ilustrada na Figura 26, que tem por objetivo atribuir um valor entre -1 à 1 de correlação linear entre 2 variáveis. Quanto mais próximo das extremidades o valor estiver, mais forte é a correlação, sendo esta positiva ou negativa.

Conforme é possível observar na Figura 26, a correlação mais forte ocorre entre as variáveis IPCA e Renda média. Quanto maior o IPCA, menor é a Renda Média (as variáveis são consideradas anti-correlacionadas).

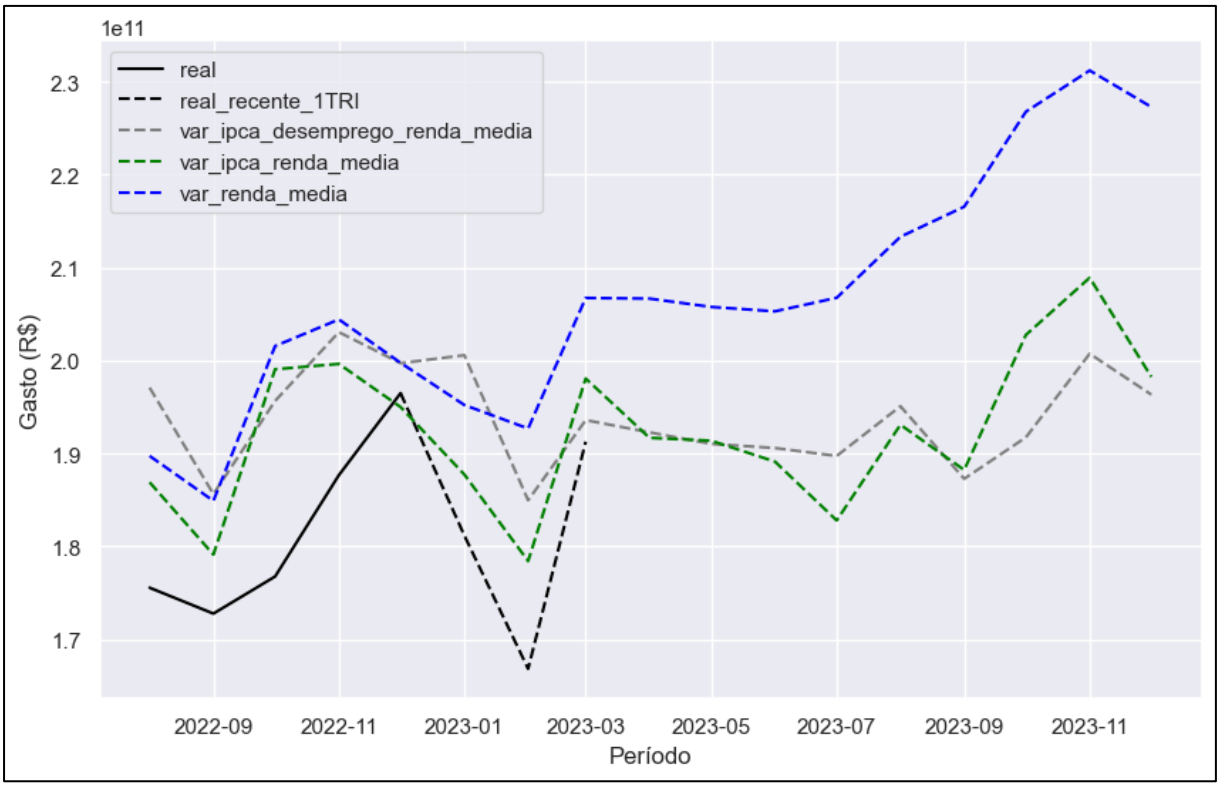

Figura 25 - Cenários previstos pelos modelos VAR multivariáveis em base normalizada

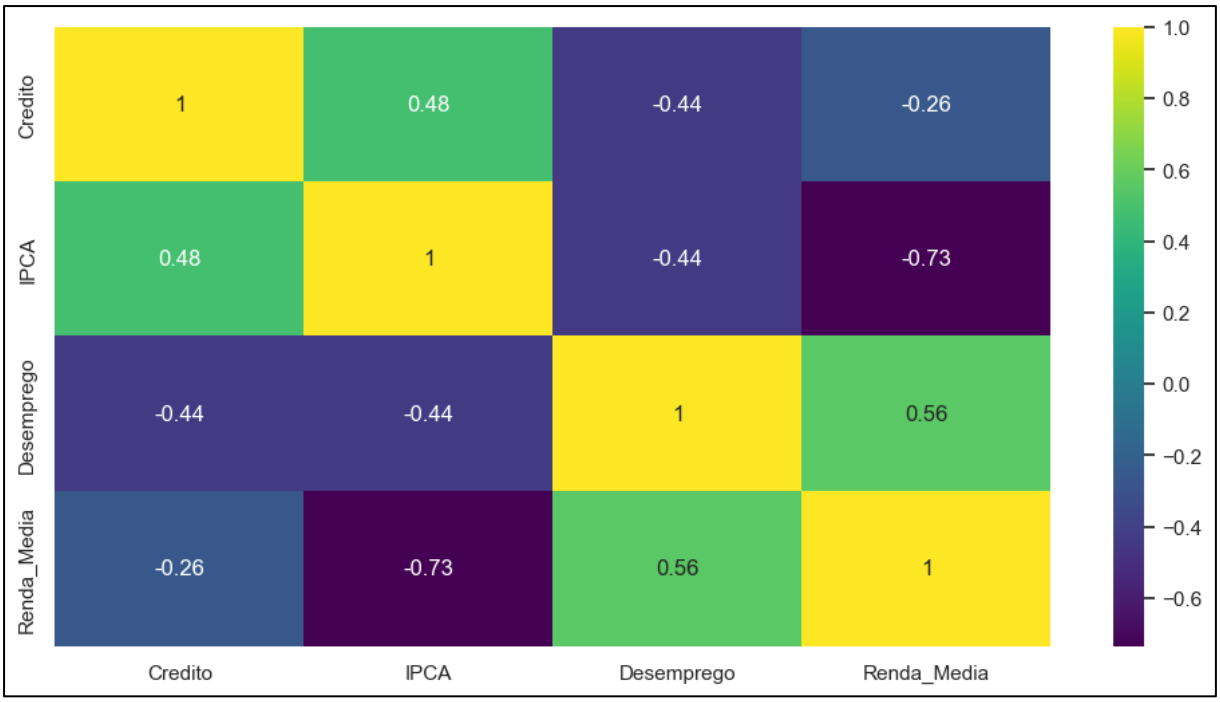

Figura 26 - Matriz de correlação de Pearson para variáveis econométricas

Com o exposto acima, pode-se assumir que a curva do IPCA com a renda média poderá apresentar desempenho melhor a longo prazo devido sua forte correlação linear; portanto, o melhor modelo obtido com a análise multivariada foi o gerado pelo algoritmo VAR com o conjunto de variáveis crédito normalizado, IPCA e renda média.

5. CONCLUSÃO

O cenário pandêmico da COVID-19 apresentou grandes desafios não apenas para a área de saúde, mas também para a economia mundial. Nos experimentos realizados neste trabalho, mesmo realizando um ajuste nos valores do ano de 2020, notou-se uma forte alteração de comportamento em 2021 e 2022. É difícil dizer se esta mudança é temporária ou se precisará de mais tempo para reduzir o impacto nas séries históricas. Entretanto, ao realizar o ajuste nos dados, obteve-se maior confiabilidade na predição dos anos posteriores a 2020, evidenciados com a redução significativa da aleatoriedade da série e uma tendência linear mais definida, quando comparada à série original.

Na análise univariada, com a transformação da série não-estacionária em estacionária, através da técnica de diferenciação logarítmica, foi possível construir um modelo para prever o volume de crédito. O melhor modelo obtido foi o Sarima com um MAPE de 6,54% e sazonalidade de 12 meses. Além de ter o melhor MAPE dentre seus concorrentes, ficou evidenciado pela Figura 23 que seu ajuste foi o mais seguro e também o mais próximo da realidade.

Já na análise multivariada, os resultados apontados pelo algoritmo Prophet não foram tão conclusivos, portanto, não foram considerados para a análise. Enquanto que as observações registradas pelo VAR apresentaram melhores desempenhos. Além disso, foi possível identificar uma melhora progressiva nos resultados com a adição de variáveis econométricas. Em primeiro, a análise crédito x renda média alcançou um MAPE de 7,94%, e posteriormente, ao adicionar a este conjunto a variável IPCA, o MAPE foi reduzido para 5,97%, aproximadamente 2% de redução, sendo O melhor resultados dentre todas as previsões obtidas. Portanto, é possível concluir que introduzindo de forma separada as variáveis econométricas ao modelo, não obteve-se um resultado melhor do que o da análise univariada. Porém, quando as variáveis econométricas são trabalhadas em conjunto, o resultado do modelo pode melhorar significativamente.

REFERÊNCIAS

ABECS – ASSOCIAÇÃO BRASILEIRA DAS EMPRESAS DE CARTÃO DE CRÉDITO E SERVIÇOS. **Balanço do setor 3º trimestre de 2022 – Apresentação (11.2022)**. 10 de nov. de 2022. Disponível em: < https://abecs.org.br/apresentacoes-e-estudos>. Acesso em: 15 de jan. de 2023.

CANZIAN, Fernando. Desemprego dobra e 'inflação dos pobres' dispara 40%. **Folha de São Paulo**, São Paulo, 30 de out. de 2021. Disponível em: <https://www1.folha.uol.com.br/mercado/2021/10/desemprego-dobra-e-inflacao-dos-pobres-dispara-40-na-pandemia.shtml>. Acesso em: 20 de jan. de 2023.

GUIMARÃES, Pedro. Maioria dos brasileiros tem três ou mais cartões de crédito, diz Serasa. **CNN**, Rio de Janeiro, 11 de mai. de 2022. Disponível em: <https://www.cnnbrasil.com.br/economia/maioria-dos-brasileiros-tem-tres-ou-mais-cartoes-de-credito-diz-serasa/>. Acesso em: 20 de jan. de 2023.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA . **PNAD Contínua**. Rio de Janeiro: IBGE, 2012.

MCKINNEY, Wes. **Python para Análise de Dados: tratamento de dados com pandas, numpy e ipython**. 2. ed. São Paulo: Novatec, 2018.

MORETTIN, Pedro A.; BUSSAB, Wilton de O.. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2004.

MOURA, Karina. Ciclo de vida dos dados: CRISP-DM. **Medium**, 14 de jan. de 2019. Disponível em: <https://medium.com/@kvmoura/crisp-dm-79580b0d3ac4>. Acesso em: 20 de jan. de 2023.

MULLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning With Python**. Sebastopol: O'Reilly, 2016.

NIELSEN, Aileen. **Análise Prática de Séries Temporais: predição com estatística e aprendizado de máquina**. Rio de Janeiro: Alta Books, 2021.

PENG, Roger D.; MATSUI, Elizabeth. **The Art of Data Science: a guide for anyone who works with data**. Baltimore: LLC, 2016.

RIOS, Ricardo A. **Modelagem de séries temporais por meio da decomposição e análise de influências estocásticas e determinísticas**. Tese (Doutorado em Ciência da Computação e Matemática) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Carlos. São Paulo, 2013.

WHEELAN, Charles. **Estatística: o que é para que serve como funciona**. Rio de Janeiro: Zahar, 2016.

WHEELAN, Charles; FAWCETT, Tom. **Data Science para Negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Rio de Janeiro: Alta Books, 2016.