“A lesson without pain is meaningless, that’s because no one can gain without sacrifice something. But by enduring that pain and overcoming it, he shall obtain a powerful, unmatched heart”.

(Hiromu Arakawa, [Fullmetal Alchemist, Vol. 25](https://www.goodreads.com/work/quotes/13285879))

**Análise preditiva do IPCA com a divulgação dos dados fornecidos pela IBGE**

***IPCA predictive analysis with the release of data provide by IBGE***

Adriano Ferriani Madureira Pontes.

[**driferriani@gmail.com**](mailto:driferriani@gmail.com)

Centro Universitário Facens - Sorocaba, SP, Brasil

Submetido em: 08 04. de 2023.

**SUMÁRIO**

[1. INTRODUÇÃO 5](#_Toc131714276)

[2. MATERIAIS E MÉTODOS 7](#_Toc131714277)

[2.1. Contexto POF 7](#_Toc131714278)

[2.2. Base de Dados 7](#_Toc131714279)

[2.3. Modelagem dos dados 9](#_Toc131714280)

[2.4. Modelo estatístico 13](#_Toc131714281)

[2.5. Dessazonalização da séria histórica e previsão 17](#_Toc131714282)

[2.6. Métricas de avaliação 23](#_Toc131714283)

[3. CONCLUSÃO 24](#_Toc131714284)

[REFERÊNCIAS 25](#_Toc131714285)

**RESUMO**

Observa-se a Inflação do Brasil com o índice IPCA (Índice Nacional de Preços ao Consumidor Amplo) é determinante para a leitura das atividades econômicas do país, e entender qual o curso que ela está tomando. É essencial para todos, (estado/governo, empresas e famílias) poder se planejar com seus recursos. A leitura do senário macroeconômico, facilita o entendimento do impacto que gera no custo de vida das pessoas, como por exemplo as mudanças da taxa básica de juros chamada de Selic e entre outros. Para o IBGE (Instituto Brasileiro de Geografia e Estatística) sendo um órgão público tem a obrigação da leitura da Inflação do Brasil, através do SIDRA (Sistema IBGE de Recuperação Automática) onde é responsável por coletar dados dos preços de produtos dentro do nosso país. A coleta é feita dentro de regiões dos cinco estados mais importantes do país, dos preços dos produtos, onde as marcas ficam em sigilo. A partir deles será divulgada a variação dos preços dos produtos e o peso dentro da composição do IPCA. Esse trabalho tem de apresentar a coleta desses valores dos anos passados, aplicar a modelagem de dados para poder enfim utilizar modelos estatísticos de sazonalidade como a ARMA, ARIMA e SARIMA como nossa base de análise preditiva.

**Palavras chaves:** IPCA, Inflação, IBGE, Sazonalidade, Modelos de predição, ARMA, ARIMA, SARIMA.

**ABSTRACT**

It is observed the Inflation in Brazil with the IPCA index (National Index of Prices to the Ample Consumer) is determinant for the reading of the economic activities of the country, and to understand which course it is taking. It is essential for everyone (state/government, companies and families) to be able to plan with their resources. Reading the macroeconomic scenario makes it easier to understand the impact it generates on people's cost of living, such as changes in the basic interest rate called Selic, among others. For the IBGE (Brazilian Institute of Geography and Statistics) being a public body, it has the obligation to read the Inflation in Brazil, through SIDRA (IBGE Automatic Recovery System) where it is responsible for collecting data on product prices within our country. The collection is made within regions of the five most important states of the country, the prices of the products, where the brands are kept confidential. From them, the variation in product prices and the weight within the composition of the IPCA will be disclosed. This work must present the collection of these values ​​from past years, apply data modeling to finally be able to use statistical models of seasonality such as ARMA, ARIMA and SARIMA as our basis for predictive analysis.

Keywords: IPCA, Inflation, IBGE, Seasonality, ARMA, ARIMA, SARIMA.

# INTRODUÇÃO

A Inflação é um assunto em comum pelos economistas por se tratar de um efeito que gera muito impacto na vida das pessoas. O conceito da inflação estende a um aumento generalizado dos preços de bens e serviços. Essa variação dos preços dos produtos é um fator que pode ser observado de forma macroeconômica, e caso analisada, esclarece informações sobre o poder de compra das pessoas e sobre os recursos disponíveis no mercado. O indicie IPCA demonstrações das variações com causas de inflação ou um efeito ao contrário chamado de deflação, pela forma apresentada pelo IBGE, se retrata ao valor de variação dos preços coletados dos itens dos produtos. Vale lembrar que o IPCA tratado no trabalho não é o único índice que pode se avaliar a saúde econômica de um país.

Para a macroeconomia, o conceito de inflação é mensurável, traçando uma cesta de itens de produtos formando um índice como medidor. O tema na macroeconomia é muito extenso, existem diversas variáveis influentes a causa desse efeito. O foco do trabalho é a análise preditiva referente a coleta de preços do IBGE do indicador IPCA. Essas análises para o mercado de capitais se tornam de grande importância. O próprio Bacen necessita dessas avaliações de forma precisa para poder tomar atitudes no mercado econômico do país.

A IPCA é composta por uma cesta dos preços dos produtos e serviços, onde atualmente é coletado por 377 itens, o IBGE separou cada um dentro dos nove grupos. Para cada um desses itens, a coleta é realizada e divulgada com a variação do preço e o peso composto dentro da carteira do consumidor. Para se chegar ao resultado do Índice Geral (IPCA) o cálculo se dá pela soma dos produtos das variações e pesos de cada item dividido pela soma total dos pesos de cada item. (InfoMoney, 2022,ibid).

Este estudo em particular, seu objetivo é analisar a variação dos preços nos próximos meses. Se o IBGE divulgasse o nome dos 377 itens coletados, o formato da pesquisa mudaria e a coleta dos preços dos produtos como forma de entrada tornaria os dados mais precisos na hora de almejar um resultado.

Para realizar a previsão da Inflação do IPCA, é necessária antes coletar os dados disponíveis pelo IBGE. Essa tarefa se torna árduo quando se trata em coletar esse tipo de informação, e montar uma série temporal completa.

No trabalho será mais bem detalhado dificuldades e obstáculos enfrentados na coleta de dados do IPCA, bem como apresentar justificativas para as decisões tomadas na criação de uma série temporal, obtendo valores coerente e adequada com a modelagem de dados, adquirindo valores no início dos anos de 1998.

Dessa forma, serão abordados aspectos relevantes da metodologia de coleta de dados utilizada pelo IBGE, bem como os desafios enfrentados na obtenção de dados precisos e confiáveis. Além de demonstração das estratégias utilizadas para criar uma linha temporal completa e coerente, a partir da qual será possível realizar nossos modelos de estatística.

# MATERIAIS E MÉTODOS

# Contexto POF

A Pesquisa de Orçamentos Familiares (POF) é uma investigação periódica conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE), a qual tem por objetivo atualizar e monitorar o consumo das famílias brasileiras. A POF influencia a estrutura dos itens de pesquisa e coleta de preços dos produtos, sendo a base para a determinação do Índice de Preços ao Consumidor Amplo (IPCA).

A mais recente POF compreende uma lista de 377 produtos, cuja composição é essencial para a mensuração do IPCA. Cada um desses itens é identificado por um valor único de ID, como é o caso do produto "Arroz", que é identificado pelo ID 1101002.

No entanto, como as preferências de consumo dos indivíduos mudam com o tempo, a POF busca manter atualizada a composição dos produtos que compõem a cesta de consumo dos brasileiros. Por exemplo, a TV com antena era popular em um passado recente, mas atualmente as smart TVs têm ganhado espaço no mercado. A POF tem como objetivo acompanhar essas mudanças nas preferências de consumo e atualizar a composição da cesta de produtos de acordo com as mudanças na carteira de consumo dos cidadãos.

É importante destacar que, ao longo dos anos, houve diversas edições da POF, o que significa que podem existir produtos que já não são mais consumidos pelas famílias, bem como itens que surgiram recentemente. Portanto, é necessário que haja constante atualização da lista de produtos pesquisados.

O IBGE disponibiliza um vídeo em seu portal IBGE Educa (2015) explicando o significado e a importância da POF. A análise desses dados é fundamental para a mensuração da inflação e para a tomada de decisões econômicas.

# Base de Dados

O Instituto Brasileiro de Geografia e Estatística (IBGE) dispõe de duas fontes de dados distintas, cada uma apresentando suas particularidades em relação aos valores contidos. A primeira delas é proveniente diretamente do IBGE SIDRA, onde mensalmente é divulgado o Índice Nacional de Preços ao Consumidor Amplo (IPCA), contendo dados armazenados exclusivos da Pesquisa de Orçamentos Familiares (POF) atual no período de 2019 até o presente momento. É possível acessar os dados por meio de uma API, que fornece resultados referentes à variação e ao peso que compõem o IPCA de cada mês, disponível no endereço eletrônico https://sidra.ibge.gov.br/home/ipca/brasil.

Para este trabalho, foi adotada como estratégia a utilização dos dados provenientes de uma série histórica mais ampla, disponibilizada pelo IBGE Agregado, que pode ser acessada por meio do endereço https://servicodados.ibge.gov.br/api/docs/agregados?versao=3. Diversas APIs são disponibilizadas, cada uma contendo informações referentes às POFs em períodos distintos. No entanto, a dificuldade encontrada nessa fonte de dados reside na necessidade de dessazonalizar para cada POF e tratar os dados para cada item que fez parte da cesta do IPCA.

Na obtenção dos dados, foram utilizadas diversas APIs, conforme organizado na Tabela 1. O conteúdo disponibilizado por elas contempla a variação e o peso dos itens, categorias, subgrupos, grupos e o IPCA (nomeado como Índice Geral).

Tabela 1 – APIs do IBGE Agregado

|  |  |
| --- | --- |
| **POF Base de dados** | |
| Período tempo | API da base |
| 2020/jan até 2022/nov | [https://servicodados.ibge.gov.br/api/v3/agregados/7061/periodos/202001|202002|202003|202004|202005|202006|202007|202008|202009|202010|202011|202012|202101|202102|202103|202104|202105|202106|202107|202108|202109|202110|202111|202112|202201|202202|202203|202204|202205|202206|202207|202208|202209|202210|202211/variaveis/306|309?localidades=N1[all]&classificacao=315[all]](https://servicodados.ibge.gov.br/api/v3/agregados/7061/periodos/202001|202002|202003|202004|202005|202006|202007|202008|202009|202010|202011|202012|202101|202102|202103|202104|202105|202106|202107|202108|202109|202110|202111|202112|202201|202202|202203|202204|202205|202206|202207|202208|202209|202210|202211/variaveis/306|309?localidades=N1%5ball%5d&classificacao=315%5ball%5d) |
| 2012/jan até 2019/dez | https://servicodados.ibge.gov.br/api/v3/agregados/1420/periodos/201201|201202|201203|201204|201205|201206|201207|201208|201209|201210|201211|201212|201301|201302|201303|201304|201305|201306|201307|201308|201309|201310|201311|201312|201401|201402|201403|201404|201405|201406|201407|201408|201409|201410|201411|201412|201501|201502|201503|201504|201505|201506|201507|201508|201509|201510|201511|201512|201601|201602|201603|201604|201605|201606|201607|201608|201609|201610|201611|201612|201701|201702|201703|201704|201705|201706|201707|201708|201709|201710|201711|201712|201801|201802|201803|201804|201805|201806|201807|201808|201809|201810|201811|201812|201901|201902|201903|201904|201905|201906|201907|201908|201909|201910|201911|201912/variaveis/306|309?localidades=N1[all]&classificacao=315[all] |
| 2006/jul até 2011/dez | [https://servicodados.ibge.gov.br/api/v3/agregados/2942/periodos/200607|200608|200609|200610|200611|200612|200701|200702|200703|200704|200705|200706|200707|200708|200709|200710|200711|200712|200801|200802|200803|200804|200805|200806|200807|200808|200809|200810|200811|200812|200901|200902|200903|200904|200905|200906|200907|200908|200909|200910|200911|200912|201001|201002|201003|201004|201005|201006|201007|201008|201009|201010|201011|201012|201101|201102|201103|201104|201105|201106|201107|201108|201109|201110|201111|201112/variaveis/306|309?localidades=N1[all]&classificacao=315[all]](https://servicodados.ibge.gov.br/api/v3/agregados/2942/periodos/200607|200608|200609|200610|200611|200612|200701|200702|200703|200704|200705|200706|200707|200708|200709|200710|200711|200712|200801|200802|200803|200804|200805|200806|200807|200808|200809|200810|200811|200812|200901|200902|200903|200904|200905|200906|200907|200908|200909|200910|200911|200912|201001|201002|201003|201004|201005|201006|201007|201008|201009|201010|201011|201012|201101|201102|201103|201104|201105|201106|201107|201108|201109|201110|201111|201112/variaveis/306|309?localidades=N1%5ball%5d&classificacao=315%5ball%5d) |
| 1999/ago até 2006/jun | https://servicodados.ibge.gov.br/api/v3/agregados/661/periodos/199908|199909|199910|199911|199912|200001|200002|200003|200004|200005|200006|200007|200008|200009|200010|200011|200012|200101|200102|200103|200104|200105|200106|200107|200108|200109|200110|200111|200112|200201|200202|200203|200204|200205|200206|200207|200208|200209|200210|200211|200212|200301|200302|200303|200304|200305|200306|200307|200308|200309|200310|200311|200312|200401|200402|200403|200404|200405|200406|200407|200408|200409|200410|200411|200412|200501|200502|200503|200504|200505|200506|200507|200508|200509|200510|200511|200512|200601|200602|200603|200604|200605|200606/variaveis/306?localidades=N1[all]&classificacao=315[all] |
| 1998/jan até 1999/jul | [https://servicodados.ibge.gov.br/api/v3/agregados/645/periodos/199801|199802|199803|199804|199805|199806|199807|199808|199809|199810|199811|199812|199901|199902|199903|199904|199905|199906|199907/variaveis/306?localidades=N1[all]&classificacao=72[all]](https://servicodados.ibge.gov.br/api/v3/agregados/645/periodos/199801|199802|199803|199804|199805|199806|199807|199808|199809|199810|199811|199812|199901|199902|199903|199904|199905|199906|199907/variaveis/306?localidades=N1%5ball%5d&classificacao=72%5ball%5d) |

# Modelagem dos dados

A modelagem de dados é uma atividade que demanda um conhecimento prévio sobre como os valores são definidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A estrutura dos dados é organizada em colunas, sendo que cada coluna representa uma data e é demarcada por um identificador (ID) associado ao nome do assunto tratado e suas variações.

Importante mencionar que a classificação em categorias, subgrupos e grupos foi estabelecida pelo IBGE, mas não os seus respectivos nomes. Neste contexto, optou-se por utilizar essa nomenclatura com o objetivo de facilitar a comunicação e compreensão dos dados.

Os grupos abrangem a maior parte dos itens, seguidos pelos subgrupos e, por fim, as categorias, que estão mais próximas dos itens. O IBGE utiliza um sistema de identificação, em que o ID possui um dígito para grupos, dois para subgrupos, quatro para categorias e sete para itens.

Para a análise desses dados, é necessário criar colunas que distinguem o número do ID, uma coluna com o nome do produto e outra para demarcar em qual posição da hierarquia as variações se encontram, com o nome de “groupSeparater”. A Figura 1 apresenta um exemplo de tabela em formato data frame com essa estrutura.

O primeiro passo para a análise é coletar as tabelas de cada API e concatená-las em uma ordem temporal, de janeiro de 1998 a novembro de 2022. Em seguida, as colunas devem ser separadas em três categorias com valores categóricos.

Tabela 1 – dados de variação do IBGE

Tabela

Descrição gerada automaticamente

A estratégia inicial para a análise preditiva com o modelo estatístico de série temporal consiste em conhecer o dado fornecido e analisá-lo sob diferentes perspectivas, a fim de facilitar a escolha do caminho a ser tomado. Nesse sentido, é fundamental realizar a identificação de valores nulos em toda a série histórica, conforme ilustrado na imagem 2. Os valores nulos são representados por áreas escuras, enquanto as áreas amarelas indicam valores presentes. É possível observar que as cinco POFs estão claramente separadas em blocos retangulares, o que facilita a identificação de cada uma delas. Além disso, foram identificados conjuntos e itens que estão presentes desde o início de 1998 e que perduraram até os dias atuais, bem como outros itens que aparecem e desaparecem em datas mais recentes. Também foram identificados itens que desapareceram por um tempo e reapareceram na última POF, a mais atual, e valores que surgiram somente nessa última POF.

Imagem2 – Heatmap para valores nulos

Gráfico

Descrição gerada automaticamente com confiança média

A partir disso, a estratégia na tomada de decisão começa a se concretizar. Os itens mais importantes estão dentro da Pesquisa de Orçamentos Familiares (POF) mais atual e a análise será concentrada nos grupos e nos itens. A análise abordada apenas aos grupos demonstrou o melhor dos interesses para esse projeto, os nove grupos surgiram no início da série histórica e se mantiveram até os dias atuais, sem alteração de entrada e retirada de qualquer novo ID que possa criar uma surpresa. No entanto, a análise dos itens apresentou complicações, muitas barreiras e verificações para conseguir juntar a melhor série histórica para aqueles itens que demonstraram mais sentido, mesmo assim era a minoria dos itens com uma extensa série histórica coesa. Então a opção era manter o período da última POF para os 377 itens, perdendo a longa série histórica, ou apenas centralizar os esforços nos grupos e no índice geral. Como o tópico do projeto era no foco de manter um período externo de dados ao longo do tempo, optou-se pela segunda opção já que a primeira não faria o menor sentido em limitar o estudo para aquela base dessazonalizada no período curto de série temporal, sabendo que há uma outra API com dados mais atuais, sem nenhuma modificação prévia.

A análise dos grupos e a limpeza e modelagem de dados iniciaram-se com o filtro, separando apenas os valores de um dígito. Comparando a quantidade de grupos da POF atual com a série histórica passada, verificou-se uma diferença de um valor a mais para os períodos de POFs anteriores. A partir desse contexto, identificou-se esse grupo adicional, demonstrando ser o mesmo ID número 5, porém com dois nomes diferentes. Ao lidar com o Heat map na imagem 3, pode-se observar que para a POF de 1998 a 1996 junho tratava-se do mesmo ID grupo, porém com a mudança de seu nome, resultando na divisão das duas linhas e em valores nulos em uma das datas, enquanto o outro ID continha valores de variação. Como solução, propõe-se a aplicação de uma linha temporal coesa pela junção dos dois valores em um único ID 5.

Imagem 3 – Observação de valores nulos em cor escura dos dois grupos

Forma, Retângulo

Descrição gerada automaticamente

Após o tratamento do Grupo de ID 5, procedeu-se à verificação dos valores ausentes nas variáveis de Comunicação e Educação nos nove grupos. Para solucionar esse problema, optou-se por remover as colunas com valores nulos na primeira POF, em que não há dados disponíveis para a variável de peso, mas apenas para os valores de variação.

Com a obtenção da tabela alinhada às expectativas deste trabalho em relação aos valores de variação e peso de todos os nove grupos e do Índice Geral em uma série histórica que se estende de julho de 2006 a novembro de 2022, mantendo ao menos quatro das cinco POFs, é possível utilizá-la como fonte para a aplicação de métodos de análise preditiva de modelos estatísticos. Cabe destacar que o peso do IPCA é sempre igual a 100, sendo a soma dos pesos de todos os itens ou dos nove grupos. Enquanto o número de ID do IPCA será considerado como valor 0, os grupos são identificados pelos IDs de 1 a 9, correspondendo, respectivamente, a Alimentação e bebida, Habitação, Artigos de residência, Vestuário, Transporte, Saúde e cuidados pessoais, Despesas pessoais, Educação e Comunicação.

Imagem 4 – tabela resultado para a aplicação dos modelos.

Tabela, Excel

Descrição gerada automaticamente

# Modelo estatístico

Uma série temporal é entendida como um conjunto de valores organizados em uma linha do tempo, com valores referentes a um intervalo de tempo regular entre eles. A série temporal apresenta movimentos característicos que podem ser divididos em quatro tipos principais: tendência, ciclo, sazonalidade e aleatoriedade. Existem três tipos de situações que devem ser levadas em consideração na análise de séries temporais. O primeiro tipo corresponde às séries temporais com características de independência, que podem ser identificadas por meio da detecção de ruído branco. As outras duas situações envolvem a memória longa e a memória curta, que dependem do quanto os valores históricos afetam os valores futuros.

O conceito de ruído branco é fundamental para a análise e previsão de séries temporais. Identificado como uma série temporal não previsível e estacionária, o ruído branco é a série temporal desejada para aplicar os modelos de análise preditiva. A análise para identificá-lo é realizada por meio de conceitos estatísticos, como a média e o desvio padrão, que são constantes ao longo do tempo e não mudam com o tempo. Além disso, a correlação entre a série temporal e sua versão defasada não é significativa.

Para a construção deste trabalho, a análise será feita utilizando três conceitos estatísticos e uma verificação de teste de hipótese de DICKEY-FULLER. O objetivo é identificar se a série é estacionária ou não. Caso a série seja estacionária, será possível aplicar os modelos estatísticos e fazer a previsão. Caso contrário, será necessário realizar uma tratativa para torná-la compatível com o modelo.

A diferenciação com lags será aplicada para tratar as séries temporais. O formato utilizado para este trabalho é com a diferença de duas séries, uma equivalente à outra, e o lag rebatendo em uma delas com o número de meses atrasados. O resultado dessa diferença será novamente verificado pelo teste de hipóteses e pelos conceitos de avaliação estatística para identificar uma série temporal de ruído branco e estacionária.

É importante destacar o uso da média móvel na análise de séries temporais. A média móvel é definida como a soma de todos os valores dividida pelo número de observações. Neste trabalho, a média móvel será separada em dez cortes diferentes em linhas de tempo distintas para aplicá-la dentro desses períodos separados.

Da mesma forma, o desvio padrão (DP) será utilizado como métrica semelhante à média. O DP é a variabilidade ou a diferença entre cada valor e a média da distribuição. O desvio padrão é a raiz quadrada da variância, que é a média dos quadrados das diferenças entre cada valor e a média da distribuição.

Os modelos estatísticos utilizados neste estudo envolvem a análise de séries temporais com o objetivo de prever dados futuros. Para atingir esse objetivo, é necessário criar uma série temporal dessazonalizada, que consiste em realizar um ajuste sazonal ou correção de tendência para remover as variáveis que impactam a série de maneira estatisticamente significativa.

Uma série temporal é entendida com um conjunto de valores dentro de uma linha do tempo, os valores são referentes dentro de um intervalo de tempo regular entre eles. A movimentos característicos dentro da série temporal que vale a citação divididos em quatro tipos principais: tendência, ciclo, sazonalidade e aleatoriedade. Referente a análise de séries temporais a três tipos de situações que levar em conta. As séries temporais com características de independente, a qual é comumente identificada através da detecção de ruído branco. Essa característica pode ser especialmente relevante para grupos específicos, os quais serão discutidos com maior detalhe posteriormente. E as outras duas de memória longa, onde valores históricos causa impacto em períodos específicos. Memória curta, dependendo do quanto valores passados afetam os valores futuros. (SANTOS, 2019, p. 4, ibid).

Na tabela 2 haverá um melhor detalhamento para qual modelo deve ser aplicado a série temporal.

Tabela 2 – Modelos estatísticos

|  |  |
| --- | --- |
| Modelo | Parâmetro |
| AR(p) Modelo de Regressão Automática | Onde o parâmetro p é um número inteiro que define quantas séries defasadas foram utilizadas para fazer a previsão dos próximos períodos. |
| MA(q) Modelo média móvel | Onde nosso parâmetro q representa o número de erros da nossa previsão defasada ocorreu para dentro da equação a melhorar nossa previsão futura. Portanto se q = 1, o valor que levará em conta é apenas o erro do valor mais recente. Se q = 2 os últimos dois valores mais recentes serão levados em conta. |
| ARMA(p,q) | Junção dos dois modelos AR e MA. Devemos colocar o parâmetro p e o parâmetro q. |
| ARIMA(p,d,q) | Junção dos dois modelos (AR e MA) e adicionalmente a parte de integrada com a adição do I entre os dois (AR I MA). Usando o parâmetro d. Esse parâmetro é referente à ordem de diferenciais utilizada no conjunto de dados (Quantidade de lags utilizados). |
| SARIMA(p,d,r),(PDR)m | O S se refere a Sazonal. Ambos SARIMA e ARIMA são usados para prever dados sazonais. Os parâmetros p,d,r são o mesmo de P,D,R. Mas o parâmetro m é referente ao número sazonal, quantas vezes no período de um ano a sazonalidade é repetida. |

A fim de analisar a possibilidade de uma série temporal ser considerada como ruído branco, é necessário utilizar a Função de Autocorrelação (ACF). Além disso, a ACF é útil para determinar a ordem do modelo Moving Average (MA) por meio da identificação do valor "q" relacionado ao número de defasagens. Por outro lado, a Função de Autocorrelação Parcial (PACF) é utilizada para determinar o valor "p", que indica a ordem do modelo Autorregressivo (AR).

Para estabelecer os parâmetros e a definição do modelo estatístico para a realização da predição, são realizadas análises para cada grupo. A definição dos valores dos parâmetros p e q é feita com base nos resultados obtidos pela ACF e PACF, enquanto o valor do parâmetro d corresponde ao número de lags necessários para manter a série estacionária. Já o valor do parâmetro m é determinado pelo número sazonal de ocorrências ao longo do ano (Singh, 2020).

Para avaliar o desempenho da modelagem, foram utilizadas duas métricas estatísticas: a Raiz do Erro Quadrático Médio (RMSE) e a Média do Erro Absoluto (MAE). Ambas as métricas verificam a diferença média entre os valores previstos e os valores reais. Essas métricas são comumente utilizadas em modelos de previsão, como os modelos de regressão (Acharya, 2021).

É importante notar que o objetivo das métricas é obter valores próximos a zero, o que indica um desempenho satisfatório do modelo. O desempenho do modelo é avaliado por meio da quantificação do erro cometido. A métrica RMSE é mais sensível ao erro do que a métrica MAE, que apresenta uma sensibilidade moderada.

A equação 1 mostra o cálculo da MAE, que mede a média das diferenças absolutas entre os valores reais e previstos.

Equação 1 – MAE quanto ao valor y sendo o valor real e ^y o valor predito.

Forma

Descrição gerada automaticamente com confiança média

O cálculo da RMSE mede a raiz média da diferença quadrática entre o valor real com o predito. Assim deixado na Equação 2

Equação 2 – RMSE quanto ao valor y sendo o valor real e ^y o valor predito.

Forma

Descrição gerada automaticamente com confiança média

# Dessazonalização da séria histórica e previsão

Nesta seção da conclusão do trabalho de conclusão de curso, será apresentada a aplicação das metodologias na análise de nove grupos e do índice geral, por meio da criação de uma série temporal dessazonalizada para a aplicação de modelos de previsão.

A primeira etapa consiste na extração dos valores de variação da série temporal da fonte de dados, seguida pela modelagem desses valores em uma tabela padrão com duas colunas, representando o tempo e os valores de variação correspondentes. Em seguida, será exibido um gráfico da variação específica do grupo em questão. Esse procedimento será aplicado em todas as etapas subsequentes do processo de análise.

As próximas etapas seguirão a mesma estratégia, com algumas variações. Isso inclui a análise de ruído branco e série estacionária, separação dos grupos de aprendizado e teste, aplicação do modelo ao grupo de aprendizado para dessazonalizar a série temporal e aplicação do modelo ao grupo de teste.

Iniciando com o Grupo Zero ou o Índice Geral, os valores da tabela de fonte serão extraídos e inseridos em uma tabela padrão. Isso é ilustrado na Figura 5.

Imagem 5 – Série temporal dos valores de Variação do Indice Geral.

Interface gráfica do usuário, Texto, Aplicativo, chat ou mensagem de texto

Descrição gerada automaticamente

Ao abordar o procedimento de análise de Ruido Branco para a variável Índice Geral, por meio da observação da variação de séries temporais em um gráfico na Imagem 6. Inicialmente, não se observa nenhuma característica aparente na variação do IPCA. Na Imagem 7, são apresentadas as duas métricas de avaliação, ACF e PACF, utilizadas para verificar o Ruido Branco da série, em que os gráficos mostram sinais do Ruido Branco. Na Imagem 8, há dois gráficos que apresentam os valores de média e desvio padrão. Eles comparam a média móvel e o desvio padrão móvel, sendo divididos em dez partes diferentes, em que as variações móveis são ilustradas com marcadores circulares e a variação total é ilustrada com uma linha.

Imagem 6 – Variação do IPCA

Interface gráfica do usuário, Gráfico

Descrição gerada automaticamente

Imagem 7 – Resultado de ACF e PACF do IPCA

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Imagem 8 – Ilustração de Média e DP do Índice Geral

Gráfico

Descrição gerada automaticamente

Observando os gráficos com os resultados, aqueles contidos nas imagens 6 e 7, não demonstra qualquer tendencia ou sazonalidade e a uma correlação positiva. A observação a se fazer é sobre a série temporal, inclinada similar ao ruido branco, e estacionária. Para a imagem 8 grande parte dos valores em bolinha média moveis e desvio padrão moveis, estão próximos a linhas globais, porém algumas outras não, o que deixa na dúvida a utilização do modelo de dessazonalização ou a tratativa antes de aplicar o modelo estatístico. Para uma última verificação com o teste de hipótese colocada na imagem 9. O p-valor está dentro do nível de significância, portando a série é estacionária, determinante a aplicar o modelo estatístico.

imagem 9 - teste de DICKEY-FULLER para o IPCAInterface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Modelo estatístico entra, antes a separação do treino e teste demonstrados na imagem 10. A linha vermelha representa o teste que será aplicada para o modelo. Utilizando o modelo ARMA, o valor p=1 e o valor q=2, diante disso na imagem 11, pode se observar a variação da dessazonalização do IPCA em que o modelo ARMA criou em linha laranja com a comparação da série original em azul.

Imagem 10 – separação treino e teste do IPCA

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Imagem 11 – dessazonalização do IPCA

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

Para a imagem 12, na linha em laranja a previsão do modelo, criando um traço no centro da variação dos valores da linha original do IPCA.

Imagem 12 – previsão do modelo no IPCA

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Os mesmos resultados de série estacionários foram obtidos para os grupos 1,2,3,5 e 9. Onde será demonstrada a aplicação do modelo ARMA a eles, assim consequentemente aos grupos faltantes, será demonstrada a tratativa da série temporal e mais a aplicação do modelo ARIMA.

Para o Grupo 1. ARMA foi aplicado com valor de p=1 e valor de q=1, demonstrado na imagem 13. E para a imagem 14 a previsão do modelo.

Imagem 13 – dessazonalização de Alimentação e bebidas

Interface gráfica do usuário

Descrição gerada automaticamente

Imagem 14 – previsão do modelo em grupo Alimentação e bebidas

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Para o Grupo 2. ARMA foi aplicado com valor de p=1 e valor de q=1, demonstrado na imagem 15. E para a imagem 16 a previsão do modelo.

Imagem 15 - dessazonalização de Habitação

Interface gráfica do usuário, Gráfico

Descrição gerada automaticamente

Imagem 16 – previsão do modelo em grupo Habitação

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Para o Grupo 3. ARMA foi aplicado com valor de p=21 e valor de q=5, demonstrado na imagem 17. Dados os valores usados pela ACF e PACF na imagem 18, pode se observar a correlação alta dentro da vigésima primeira tempo da PACF. E para a imagem 19 a previsão do modelo para a série temporal de Artigos de residência, acompanhado o movimento do teste.

Imagem 17 - dessazonalização de Artigos de residência

Uma imagem contendo Gráfico

Descrição gerada automaticamente

Imagem 18 – ACF e PACF de Artigos de residência

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

Imagem 19 - previsão do modelo em grupo Artigos de residência

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Para o Grupo 5. ARMA foi aplicado com valor de p=16 e valor de q=18, demonstrado na imagem 21. E para a imagem 22 a previsão do modelo.

Imagem 21 - dessazonalização de Transportes

Linha do tempo

Descrição gerada automaticamente com confiança média

Imagem 22 – previsão do modelo em grupo Transportes

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Para o Grupo 9. ARMA foi aplicado com valor de p=12 e valor de q=12 e valor d=1, demonstrado na imagem 23. E para a imagem 24 a previsão do modelo.

Imagem 23 - Dessazonalização de Comunicação

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

Imagem 24 - Previsão do modelo em grupo Comunicação

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Para os grupos 4,6,7 e 8 terá a demonstração no trabalho para a tratativa da série temporal de forma a ser estacionaria, tornando se mais próximo do ruido branco.

Ao grupo 4, a série temporal com o lag de um mês foi aplicada para fazer a diferença dele com a original, obtendo a seguinte forma na imagem 25. A imagem 26 demonstra o gráfico aplicando o modelo de série temporal estacionária para Vestuário. Portanto para o modelo ARIMA o valor aplicado ao d será igual a 1, repetenando um lag. Dos valores p=1 e q=1. Assim na imagem 27 demonstra o teste do modelo com um acompanhamento da série tratada.

Imagem 25 – Traino e teste de Vestuário

Interface gráfica do usuário, Aplicativo, Word

Descrição gerada automaticamente

Imagem 26 - Dessazonalização de Vestuário

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

Imagem 27 – Previsão do modelo em grupo Vestuário

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Ao grupo 6, a série temporal com o lag de um mês foi aplicada para fazer a diferença dele com a original, obtendo a seguinte forma na imagem 28. A imagem 29 demonstra o gráfico aplicando o modelo de série temporal estacionária para Saúde e cuidados pessoais. Portanto para o modelo ARIMA o valor aplicado ao d será igual a 1, repetenando um lag. Dos valores p=1 e q=4. Assim na imagem 30 demonstra o teste do modelo com um acompanhamento da série tratada.

Imagem 28 – Traino e teste de Saúde e cuidados pessoais

Interface gráfica do usuário, Gráfico

Descrição gerada automaticamente

Imagem 29 - Dessazonalização de Saúde e cuidados pessoais

Interface gráfica do usuário

Descrição gerada automaticamente

Imagem 30 – Previsão do modelo em grupo Saúde e cuidados pessoais

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Ao grupo 7, a série temporal com o lag de um mês foi aplicada para fazer a diferença dele com a original, obtendo a seguinte forma na imagem 31. A imagem 32 demonstra o gráfico aplicando o modelo de série temporal estacionária para Despesas pessoais. Portanto para o modelo ARIMA o valor aplicado ao d será igual a 1, repetenando um lag. Dos valores p=6 e q=1. Assim na imagem 33 demonstra o teste do modelo com um acompanhamento da série tratada.

Imagem 31 – Traino e teste de Despesas pessoais

Desenho preto e branco

Descrição gerada automaticamente com confiança média

Imagem 32 – Dessazonalização de Despesas pessoais

Interface gráfica do usuário

Descrição gerada automaticamente

Imagem 33 - Previsão do modelo em grupo Despesas pessoais Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Ao grupo 8, a série temporal com o lag de um mês foi aplicada para fazer a diferença dele com a original, obtendo a seguinte forma na imagem 34. A imagem 35 demonstra o gráfico aplicando o modelo de série temporal estacionária para Educação. Portanto para o modelo ARIMA o valor aplicado ao d será igual a 1, repetenando um lag. Dos valores p=11 e q=1. Assim na imagem 36 demonstra o teste do modelo com um acompanhamento da série tratada.

Imagem 34 - Traino e teste de Educação

Antena em fundo branco

Descrição gerada automaticamente com confiança baixa

Imagem 35 – Dessazonalização de Educação

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança baixa

Imagem 36 - Previsão do modelo em grupo Educação

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

# Métricas de avaliação

Na tabela três, encontra-se os resultados obtidos do jupyter notebook aplicando as métricas.

Tabela 3 – Resultados métricas previsão

|  |  |  |
| --- | --- | --- |
|  | RMSE | MAE |
| IPCA | 0,471653369868155 | 0,471653369868155 |
| Alimentação e bebidas | 0,703779925423885 | 0,555458268197309 |
| Habitação | 1.087351022395781 | 0.7628748565552795 |
| Artigos de residência | 0,509781786291469 | 0,389688136648191 |
| Vestuário | 0,838909167188670 | 0,781248306418850 |
| Transportes | 2,417419069546060 | 1,629827609417360 |
| Saúde e cuidados pessoais | 0,496864616339888 | 0,414070720206284 |
| Despesas pessoais | 0,406088272818747 | 0,326763881015481 |
| Educação | 0,212263721515062 | 0,170129885321822 |
| Comunicação | 0,924169541253038 | 0,631055665868124 |

# CONCLUSÃO

Com base nos resultados obtidos, pode-se concluir que a aplicação de modelos estatísticos para a previsão do IPCA apresenta resultados promissores. As metodologias utilizadas permitiram uma boa análise das séries temporais dos grupos selecionados, bem como a identificação das melhores estratégias para o tratamento dos dados faltantes e estacionariedade.

É importante ressaltar que a obtenção de valores reais coletados mensalmente pelo IBGE é fundamental para a confiabilidade dos resultados. Além disso, a análise das divulgações do IPCA a cada mês e a utilização dos modelos estatísticos para previsão dos próximos meses é uma prática já adotada por instituições financeiras.

Contudo, é importante lembrar que outros fatores externos podem influenciar o comportamento do IPCA, tais como a política e a economia do país. Portanto, a análise dessas variáveis também pode ser relevante para aprimorar a precisão das previsões.

Em suma, a aplicação de modelos estatísticos para previsão do IPCA apresenta resultados satisfatórios e pode ser uma ferramenta útil para investidores e gestores financeiros. No entanto, é importante considerar outros fatores que podem influenciar o comportamento do índice e utilizar dados confiáveis para aumentar a precisão das previsões.

# REFERÊNCIAS

ACHARYA, S. What are RMSE and MAE? A Simple Guide to evaluation metrics. Towards Data Science, 2021. Disponível em: < <https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383#:~:text=Technically%2C%20RMSE%20is%20the%20Root,actual%20values%20of%20a%20variable> >. Acesso em: 1 de março de 2023.

Bex, T. How to Detect Random Walk and White Noise in Time Series Forecasting

Find out if the target is worth forecasting. Medium, 2021, Disponível em: < <https://towardsdatascience.com/how-to-detect-random-walk-and-white-noise-in-time-series-forecasting-bdb5bbd4ef81> >. Acesso em: 17 de dezembro de 2022

O que é inflação e por que ela impacta no seu bolso? Inflação é um conceito que traduz o aumento generalizado de preços de bens e serviços. Entenda tudo sobre o índice e como ele afeta os seus investimentos. InfoMoney, 2022. Disponível em: < <https://www.infomoney.com.br/guias/inflacao/#:~:text=Infla%C3%A7%C3%A3o%20%C3%A9%20um%20termo%20da,poder%20de%20compra%20da%20moeda> > . Acesso em: 10 de dezembro de 2022

Quanto o brasileiro gasta e com o quê IBGE Explica. IBGE Educa (IBGE), 2015. Disponível em: < <https://educa.ibge.gov.br/professores/educa-recursos/20584-ibge-explica-pof.html> >. Acesso em: 3 de dezembro de 2022

Radečić, R. Time Series From Scratch — White Noise and Random Walk

Part 3 of the Time Series from Scratch Series — Learn everything about white noise and random walks, and their importance in time series forecasting. Medium, 2021, Disponível em: < <https://towardsdatascience.com/time-series-from-scratch-white-noise-and-random-walk-5c96270514d3> >. Acesso em: 24 de dezembro de 2023

SANTOS, Ana Rita Gounça Neto Varanda. Forcasting inflation in Portugal by using machine learning techniques. 2019. 63 f. Trabalho de conclusão de curso (Mestrado em Matemática e Aplicações) - Faculdade de Ciências da Universidade de Lisboa, Lisboa, 2019. Disponível em: https://repositorio.ul.pt/bitstream/10451/42088/1/ulfc125904\_tm\_Ana\_Rita\_Santos.pdf. Acesso em: 29 março. 2023.

Singh, P. Time-Series Forecasting: How To Predict Future Data Using ARMA, ARIMA and SARIMA models. Medium, 2020. Disponível em: < <https://puneet166.medium.com/time-series-forecasting-how-to-predict-future-data-using-arma-arima-and-sarima-model-8bd20597cc7b#:~:text=ARIMA%20is%20a%20model%20that,used%20to%20forecast%20periods%20ahead> >. Acesso em: 30 de janeiro de 2023

Wolffenbüttel, A. O que é? Ajuste sazonal. IPEA, 2005. Disponível em: < <https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2087:catid=28> >. Acesso em: 30 de novembro de 2022