INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Fabio Freitas de Souza Filho

MODELAGEM DE DEMANDA POR PASSAGEM AÉREA POR MEIO DO PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Graduação
2023

Curso de Engenheria Civil-Aeronáutica

CDU 656.71

Fabio Freitas de Souza Filho

MODELAGEM DE DEMANDA POR PASSAGEM AÉREA POR MEIO DO PROCESSAMENTO DE LINGUAGEM NATURAL

Orientador

Prof. Dr. Prof. Dr. Marcelo Xavier Guterres (ITA)

ENGENHERIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2023

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Freitas de Souza Filho, Fabio
Modelagem de demanda por passagem aérea por meio do processamento de linguagem natural / Fabio Freitas de Souza Filho.
São José dos Campos, 2023.
54f.

Trabalho de Graduação – Curso de Engenheria Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2023. Orientador: Prof. Dr. Prof. Dr. Marcelo Xavier Guterres.

1. Transporte aéreo. 2. Demanda (Economia). 3. Processamento da linguagem natural. 4. Planejamento estratégico. 5. Transportes. I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

FREITAS DE SOUZA FILHO, Fabio. **Modelagem de demanda por passagem aérea por meio do processamento de linguagem natural**. 2023. 54f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Fabio Freitas de Souza Filho
TITULO DO TRABALHO: Modelagem de demanda por passagem aérea por meio do processamento de linguagem natural.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2023

Fabio Freitas de Souza Filho
Rua H8B, 226
12.228-461 – São José dos Campos–SP

MODELAGEM DE DEMANDA POR PASSAGEM AÉREA POR MEIO DO PROCESSAMENTO DE LINGUAGEM NATURAL

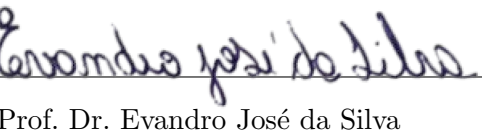
Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Fabio Freitas de Souza Filho

Autor

Prof. Dr. Marcelo Xavier Guterres (ITA)

Orientador



Prof. Dr. Evandro José da Silva
Coordenador do Curso de Engenheria Civil-Aeronáutica

São José dos Campos, 29 de novembro de 2023.

Agradecimentos

Gostaria de agradecer, inicialmente, aos meus pais que sempre deram o máximo de si para que eu conseguisse estar aqui, sou imensamente grato. Em sequência, agradeço meu irmão, que, mesmo distante, também fez parte dessa conquista.

Agora, faço agradecimentos aos amigos que fiz durante esses 5 anos, com quem sofri e sorri junto, pessoas que me apoiaram e levantaram em momentos difíceis tal como minha família fez. Estando distante geograficamente da família, foram essas pessoas que me acompanharam de perto e que eu acompanhei de perto, cito aqui alguns do meu apartamento 226- 01, 02, Biscati, Krosso e Soba- e do 215/Quarto do Pi, que sempre estiveram próximos também- Anãozin, Caio, John, Luca, Manaus, Merenda, Regis, Renan, Grilim, Cont, Maranhão e Kaysin.

Ao longo da faculdade amizades novas foram feitas, sejam por motivos de mudnça de Fund para Prof, ou por outros motivos. Alguns desses novos amigos também fizeram parte dessa história - Ana, Blaster, Cometa, Digão, Febem, Grover, Helber, Katchau, Malfoy, P+, Puto e Rasga.

Por fim, o que de fato é remanescente, são os amigos que fazemos ao longo da percurso.

"Acredito que a vida é feita para dar errado,
cabe a nós sermos inquietos quanto a isso."
— Fabio Freitas

Resumo

Este trabalho de conclusão de curso investiga a integração de variáveis geradas pelo modelo de Processamento de Linguagem Natural (PLN) Google BERT em um modelo econométrico focado na previsão de demanda por passagens aéreas. Além disso, o modelo incluiu indicadores macroeconômicos, com ênfase no Produto Interno Bruto (PIB). A pesquisa iniciou-se com a coleta de notícias relevantes ao setor aéreo por meio de técnicas de web scraping, com o objetivo de criar um banco de dados para análise pelo modelo BERT.

O propósito principal era examinar se as informações extraídas das notícias, quando convertidas em variáveis pelo BERT, poderiam enriquecer as previsões do modelo econométrico. No entanto, os resultados obtidos indicaram que, apesar da metodologia inovadora e da integração de dados não estruturados, as variáveis derivadas do BERT não apresentaram significância estatística para o modelo. Isto sugere que, no contexto específico deste estudo, as nuances linguísticas e sentimentais das notícias não tiveram impacto mensurável na demanda por passagens aéreas, quando comparadas com variáveis tradicionais como o PIB.

Este achado proporciona insights valiosos para a área de modelagem econométrica, destacando a importância de avaliar a relevância e o impacto de diferentes tipos de dados. A pesquisa realça o desafio de integrar dados de PLN em modelos econométricos e sugere a necessidade de mais estudos para explorar as condições sob as quais esses dados podem ser significativos. Este estudo contribui para o corpo de conhecimento em economia e PLN, fornecendo uma base para futuras investigações sobre a aplicabilidade de técnicas de PLN em análises econômicas.

Abstract

This thesis investigates the integration of variables generated by the Google BERT Natural Language Processing (NLP) model into an econometric model focused on forecasting air travel demand. The model also incorporates macroeconomic indicators, with an emphasis on Gross Domestic Product (GDP). The research began with the collection of relevant airline industry news through web scraping techniques, aiming to create a database for analysis by the BERT model.

The main purpose was to examine whether the information extracted from the news, when converted into variables by BERT, could enhance the forecasts of the econometric model. However, the findings indicated that despite the innovative methodology and the integration of unstructured data, the BERT-derived variables did not show statistical significance in the model. This suggests that, in the specific context of this study, the linguistic nuances and sentiments of the news did not have a measurable impact on air travel demand when compared to traditional variables like GDP.

This discovery provides valuable insights into the field of econometric modeling, highlighting the importance of assessing the relevance and impact of different types of data. The research underscores the challenge of integrating NLP data into econometric models and suggests the need for further studies to explore under what conditions such data might be significant. This study contributes to the body of knowledge in economics and NLP, providing a foundation for future investigations into the applicability of NLP techniques in economic analyses.

Lista de Figuras

Lista de Abreviaturas e Siglas

PLN	Processamento de Linguagem Natural
PIB	Produto Interno Bruto
BERT	Bidirectional Encoder Representations from Transformers
ADF	Augmented Dickey-Fuller
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function

Sumário

1 Introdução

As operações em aeroportos são estudadas de maneira profunda e com o intuito de se melhorar a eficência há bastante tempo. Esses estudos se demostram cada vez mais necessários com o aumento de demanda por passagem aérea ao longo dos anos, o que demonstra uma participação cada vez maior das viagens aéreas como meio de transporte para a sociedade, dado que o volume de passageiros desse tipo de viagem aumentou em uma taxa maior que a população mundial de 2010 a 2019.

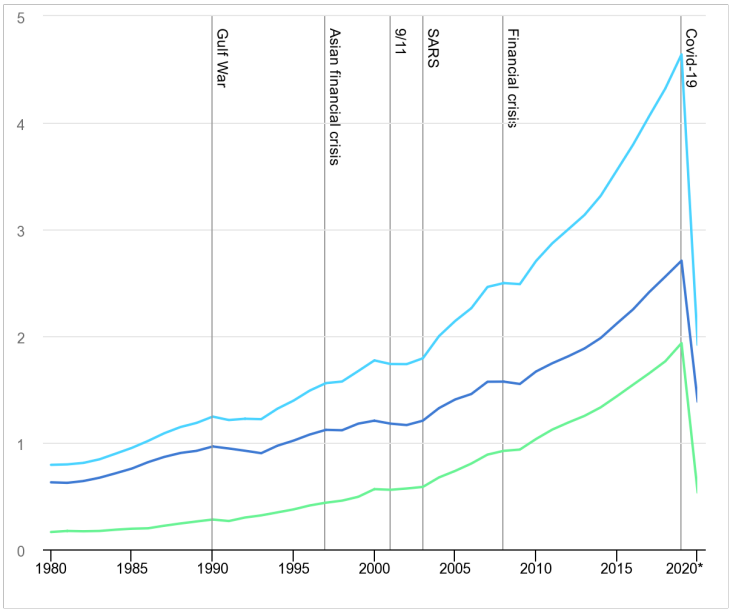


FIGURA 1.1 – Volume de passageiros global segundo a IEA.

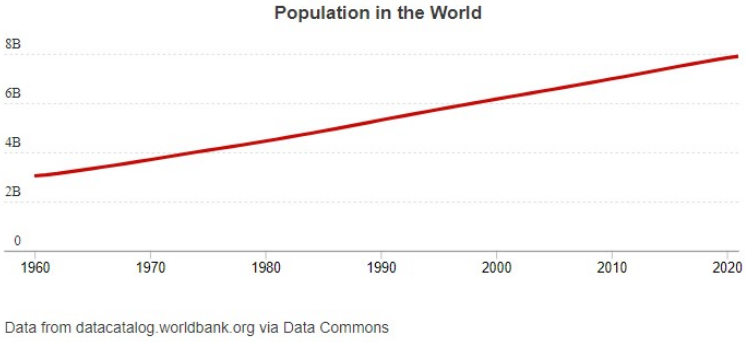


FIGURA 1.2 – População mundial segundo o World Bank.

Contra esse fato de crescimento de viagens aéreas, no período da pandemia houve uma grande queda dessa demanda, no entanto, o que se estuda hoje é que há um retorno dessa demanda por passagem aérea aos níveis de 2019 e, em países mais emergentes e na China, esse retorno é quase imediato. Portanto, existe também uma necessidade por melhora nas operações e nas capacidades dessas viagens aéreas ao longo dos anos.

Conforme essa ideia de necessidade de melhoria nas operações de aeroportos, um fator muito importante é ter uma boa projeção de passagens aéreas para dado período de tempo seja no contexto geral aeroporto ou no contexto empresarial das companhias aéreas. Com essa projeção bem feita, o aeroporto ou a empresa poderia ser se preparar ou se aperfeiçoar para abordar determinado crescimendo ou queda nas suas operações.

A previsão de demanda por passagem aérea utilizando teorias econométricas ou de machine learning é algo já explorado, inclusive com análises específicas para regiões, considerando variáveis econômicas como preço do petróleo, PIB ou a própria série temporal.

1.1 Problema de Pesquisa

Considerando, então, a alta taxa de crescimento de volume de passageiros, tem-se a necessidade de previsão de demanda por viagens aéreas para uma possível otimização das operações aeroportuárias.

Dessa forma, uma possível variável a se considerar para a descrição dessa demanda podem ser notícias relacionadas a viagens aéreas, ou que descrevam cenários econômicos. A incorporação de notícias relacionadas a viagens aéreas e cenários econômicos nas previsões de demanda é uma possível síntese sobre as tendências e padrões que influenciam o comportamento dos passageiros. Ao analisar notícias específicas do setor de viagens aéreas, é possível obter informações sobre eventos, como abertura de novas rotas, fusões entre companhias aéreas, lançamento de programas de fidelidade e mudanças nas políticas de transporte, que podem afetar diretamente a demanda por passagens aéreas.

Da mesma forma, a análise de notícias que descrevem cenários econômicos, como indicadores macroeconômicos e políticas governamentais relacionadas a viagens e turismo, pode fornecer uma visão mais ampla das condições socioeconômicas que impactam a demanda por viagens aéreas.

Além dessa descrição de cenários por notícias, métricas mais clássicas como PIB, dentre outras, também devem ser consideradas para a modelagem da demanda por passagem aérea.

1.2 Objetivo Geral

O presente estudo tem como objetivo principal a modelagem da demanda por passagens aéreas por meio da aplicação de métodos de machine learning para a análise de sentimento em notícias que contenham palavras-chave ou ”tags”específicas, além da utilização de métodos estatísticos baseados em métricas econômicas tradicionais. O propósito do modelo desenvolvido é fornecer uma base sólida para a realização de projeções da demanda por passagens aéreas no contexto brasileiro. A abordagem conjunta de machine learning e análise de sentimento visa capturar as nuances das notícias relacionadas ao setor de viagens aéreas, enquanto as métricas econômicas clássicas proporcionam uma compreensão abrangente dos fatores macroeconômicos que influenciam a demanda.

1.3 Objetivos Específicos

- Fazer o webscraping de notícias para coletar a base de dados de textos, com o intuito de se fazer a análise de sentimento.
- Coletar dados de métricas econômicas e de volume de passageiros para a modelagem e previsão.
- Utilização do modelo Google BERT, que é um modelo de processamento de lingugem natural, para a compreensão do contexto e da semântica dos textos de notícias, então o modelo deve entender relações complexas entre as palavras e retornar alguma métrica que será usada como input para o modelo de previsão de demanda.
- Juntar as variáveis econômicas e a análise de sentimento e utilizá-las como variáveis exógenas a um modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) para a previsão de demanda por passagem aérea.
- Estudar os resultados e as métricas de avaliações dos modelos, assim como entender quais variáveis fazem sentido ainda serem utilizadas e se a descrição de cenários por notícias é de fato relevante para a previsão de uma demanda futura por vôos.

1.4 Justificativa

Este trabalho é motivado pela tentativa de aprimorar a capacidade de previsão da demanda por passagens aéreas, mediante a inclusão de uma variável oriunda das notícias coletadas em um portal brasileiro. A busca é direcionada à obtenção de informações que retratem tanto o panorama econômico vigente quanto a percepção dessas circunstâncias pelos consumidores de serviços de transporte aéreo.

Ao incorporar essa variável proveniente das notícias, busca-se enriquecer o modelo de previsão existente, conferindo-lhe uma maior sensibilidade em relação às flutuações da demanda no contexto da aviação comercial. Dessa maneira, almeja-se obter uma visão mais abrangente e precisa das condições econômicas e do sentimento dos compradores de passagens aéreas, com a finalidade de aprimorar a tomada de decisões estratégicas no setor.

1.5 Organização do trabalho

A organização deste trabalho, ou seja, a forma como são apresentados os capítulos e suas descrições está explicitada a seguir:

1. *Introdução*, esse capítulo consiste em mostrar o contexto para os motivos dessa pesquisa.
2. *Revisão Bibliográfica*, após a introdução dos motivos dessa pesquisa, demonstra-se os conceitos utilziados neste trabalho.
3. *Metodologia*, nesse capítulo é apresentada os métodos que foram utilizados os conceitos anteriormente apresentados.
4. *Resultados*, o objetivo desse capítulo é apresentar os resultados obtidos a partir dos modelos utilizados.
5. *Conclusão*, com os resultados apresentados, deve-se levantar discussões e possíveis conclusões. Assim como comparar com o que era esperado.

2 Revisão Bibliográfica

2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que combina técnicas de linguística computacional, inteligência artificial e aprendizado de máquina com o objetivo de permitir que computadores entendam, interpretem e gerem linguagem humana. Essa disciplina, abordada em detalhes na obra de Eisenstein (EISENSTEIN, 2019), é essencial para o desenvolvimento de sistemas capazes de interagir e compreender a linguagem natural em diferentes aplicações, como tradução automática, chatbots, análise de sentimentos, entre outros.

O PLN envolve a construção de modelos computacionais que se baseiam em teorias linguísticas para analisar e processar textos escritos ou falados. Uma das principais tarefas do PLN é a análise sintática, que busca identificar a estrutura gramatical de uma sentença. Esse processo pode ser realizado por meio de técnicas como a análise de dependência, que busca estabelecer as relações entre as palavras de uma sentença, e a análise de árvore de constituintes, que representa a estrutura hierárquica de uma sentença por meio de uma árvore.

Outra tarefa importante no PLN, discutida em (EISENSTEIN, 2019), é a análise semântica, que visa extrair o significado das sentenças. Essa análise pode envolver, por exemplo, a classificação de sentimentos de uma sentença. Para realizar essa tarefa, são utilizadas técnicas como a representação vetorial de palavras, que mapeia palavras para vetores numéricos de alta dimensionalidade, permitindo a captura de relações semânticas entre elas.

Logo, apresentam-se como técnicas computacionais alguns algoritmos de aprendizado de máquina, como as redes neurais artificiais, que são capazes de aprender padrões e realizar tarefas de processamento de linguagem natural. Um exemplo é a utilização de redes neurais recorrentes, como as redes LSTM (Long Short-Term Memory), que têm a capacidade de capturar informações de contexto ao processar sequências de palavras.

Outro avanço importante no PLN é o uso de modelos de linguagem pré-treinados, como

o Transformer, que captura relações de dependência entre palavras e obteve resultados significativos em várias tarefas de PLN. Além disso, o desenvolvimento de assistentes virtuais inteligentes, como a Siri da Apple, a Alexa da Amazon e o Google Assistant também são produtos dos avanços de PLN, esses assistentes utilizam técnicas avançadas de PLN para compreender os comandos de voz dos usuários e fornecer respostas relevantes e contextuais.

Em síntese, o Processamento de Linguagem Natural é uma área de pesquisa e desenvolvimento em rápido crescimento, impulsionada pelos avanços em inteligência artificial e aprendizado de máquina. Com técnicas cada vez mais sofisticadas e o uso de grandes conjuntos de dados, o PLN, conforme explorado por Eisenstein (EISENSTEIN, 2019), tem o potencial de revolucionar a forma com que textos são interpretados por máquinas de maneira automática.

2.2 Modelos SARIMA

Modelos SARIMA (Seasonal Autoregressive Integrated Moving Average) são utilizados na análise de séries temporais para modelagem de séries históricas em que possuem padrões sazonais. Esses modelos são uma extensão dos modelos ARIMA (Autoregressive Integrated Moving Average) e fornecem uma formulação para considerar tanto os componentes autoregressivos quanto os componentes de média móvel em séries temporais com sazonalidade. Como discutido por Box e Jenkins (BOX; JENKINS, 2015), a composição desses modelos é feita por três fatores principais que seriam a parte autoregressiva (AR), a parte de média móvel (MA) e a parte de diferenciação (I). O componente AR modela a relação entre os valores passados da série temporal e o valor atual, enquanto o componente MA modela a relação entre os erros passados e o valor atual. A parte de diferenciação é responsável por tornar a série temporal estacionária, removendo tendências e sazonalidades.

No intuito de adicionar o efeito da sazonalidade, os modelos SARIMA também incluem um componente de diferenças sazonais (S), o qual é responsável por modelar a diferença entre os valores observados em uma determinada temporada e os valores observados na mesma temporada em períodos anteriores. Então, é possível capturar padrões sazonais e ajustar a série temporal para remover a sazonalidade.

O modelo SARIMA pode ser descrito da forma simplificada: SARIMA(p, d, q)(P, D, Q, S).

- p: ordem do componente autoregressivo.
- d: ordem do componente de diferenciação.

- q: ordem do componente de média móvel.
- P: ordem do componente autoregressivo sazonal.
- D: ordem do componente de diferenciação sazonal.
- Q: ordem do componente de média móvel sazonal.
- S: período da sazonalidade.

De maneira matemática o modelo SARIMA pode ser escrito da seguinte forma:

$$\phi_p(B)(1 - \Phi_P(B^s))(1 - B)^d(1 - B^s)^D X_t = \theta_q(B)(1 - \Theta_Q(B^s))\epsilon_t$$

- $\phi_p(B)$ é o operador autoregressivo de ordem p.
- $\Phi_P(B^s)$ é o operador autoregressivo sazonal de ordem P com período de sazonalidade s.
- $(1 - B)^d$ é o operador de diferenciação de ordem d.
- $(1 - B^s)^D$ é o operador de diferenciação sazonal de ordem D com período de sazonalidade s.
- X_t é a série temporal de interesse.
- $\theta_q(B)$ é o operador de média móvel de ordem q.
- $\Theta_Q(B^s)$ é o operador de média móvel sazonal de ordem Q com período de sazonalidade s.
- ϵ_t é o termo de erro, assumido como uma sequência de ruído branco.

Essa formulação matemática descreve a relação entre os valores passados da série temporal e os termos autoregressivos, assim como a influência dos termos de média móvel e os erros aleatórios. Os operadores B e B^s representam os operadores de defasagem simples e sazonal, respectivamente.

2.3 Modelos SARIMAX

Em um modelo econométrico existem variáveis endógenas, que são determinadas e influenciadas pelo próprio sistema em análise, e existem variáveis exógenas que podem

ser consideradas, essas variáveis podem influenciar as variáveis exógenas do sistema em análise ao descrever cenários.

Assim, para a consideração dessas variáveis exógenas em um modelo SARIMA, utiliza-se uma amplificação que seriam os modelos SARIMAX. Assim, em um modelo SARIMAX busca-se incorporar relações das variáveis exógenas, que podem ser econômicas, por exemplo, com a série temporal que está sendo estudada. Essas variáveis são consideradas como informações externas que podem ajudar a melhorar a precisão das previsões e a compreensão dos padrões da série temporal, como detalhado por Shumway e Stoffer (SHUMWAY; STOFFER, 2017).

Seguindo essa linha de raciocínio, a inclusão das variáveis exógenas ocorre adicionando termos multiplicativos à equação do modelo SARIMA, em que a influência dessas variáveis é ponderada pelos coeficientes correspondentes. Esses coeficientes, juntamente com os parâmetros do modelo SARIMA, são estimados a partir dos dados históricos da série temporal e das variáveis exógenas disponíveis.

Para a exemplificação e formulação matemática do modelo SARIMAX, tem-se o fator multiplicativo adicional das variáveis exógenas exibido da seguinte forma:

$$\phi_p(B)(1-\Phi_P(B^s))(1-B)^d(1-B^s)^D X_t = \theta_q(B)(1-\Theta_Q(B^s))\epsilon_t + \beta X_t^{(e)}$$

- $\phi_p(B)$, $\Phi_P(B^s)$, $\theta_q(B)$ e $\Theta_Q(B^s)$ são os operadores autoregressivos e de média móvel, tanto para os componentes não sazonais quanto sazonais do modelo SARIMA.
- $(1-B)^d$ e $(1-B^s)^D$ são os operadores de diferenciação de ordem d e D, respectivamente, para remover tendências e sazonalidades da série temporal.
- X_t é a série temporal de interesse.
- ϵ_t é o termo de erro, assumido como uma sequência de ruído branco.
- β é o coeficiente que representa o impacto da variável exógena $X_t^{(e)}$ no modelo SARIMAX.

A inclusão da variável exógena é representada pela adição do termo $\beta X_t^{(e)}$ à equação do modelo SARIMA. Esse termo pondera a influência da variável exógena ($X_t^{(e)}$) nos valores da série temporal. O coeficiente β quantifica o impacto dessa variável exógena na série temporal.

2.4 Modelos de previsão de demanda

Um dos estudos relevantes na análise da relação entre a demanda por transporte aéreo e o desenvolvimento econômico é conduzido por (TOLCHA, 2020). Em seu artigo *"The Impact of Economic Development on Domestic Air Traffic Demand"*, Tolcha et al. (2020) investigam essa dinâmica em seis países da África Subsaariana durante o período de 1981 a 2018. Utilizando modelos de correção de erro vetorial e autoregressão vetorial, o estudo descobre relações causais heterogêneas, específicas a cada contexto nacional. Notavelmente, foi observado que, em países como a África do Sul, Nigéria e Quênia, o desenvolvimento econômico direciona a demanda por transporte aéreo. Por outro lado, na Etiópia, o aumento na demanda por transporte aéreo é que parece fomentar o desenvolvimento econômico. Contudo, em Senegal e Angola, a relação entre as variáveis é fraca, não permitindo uma conclusão sobre a direção da causalidade. O estudo também revela que a demanda por transporte de passageiros é mais sensível às mudanças econômicas do que o volume de carga aérea, indicando variações na elasticidade da demanda por transporte aéreo em relação ao desenvolvimento econômico entre os países estudados.

Seguindo ainda a ideia, no desenvolvimento de modelos de previsão de demanda por transporte aéreo, é fundamental entender a interação entre este e a atividade econômica. O estudo de Ishutkina (ISHUTKINA, 2009), *"Analysis of the Interaction Between Air Transportation and Economic Activity: A Worldwide Perspective"*, aborda essa complexa relação, destacando como a infraestrutura aérea e o tráfego de passageiros e cargas se entrelaçam com o crescimento econômico. Ishutkina utiliza uma metodologia exploratória, combinando revisão da literatura e análise de dados de vários países, para construir um modelo de feedback que ilustra a influência recíproca entre o desenvolvimento econômico e o transporte aéreo. O estudo aponta para a importância da expansão da infraestrutura aérea e o aumento da frequência de voos como fatores-chave no estímulo ao crescimento econômico, especialmente em regiões em desenvolvimento. Estas descobertas são vitais para a formulação de modelos de previsão de demanda que consideram não apenas os padrões de viagens aéreas, mas também a dinâmica econômica subjacente, fornecendo insights importantes para planejadores e formuladores de políticas no setor de aviação.

3 Metodologia

3.1 Organização da metodologia

Segue abaixo o diagrama que representa a ordem dos tratamentos e dos modelos.

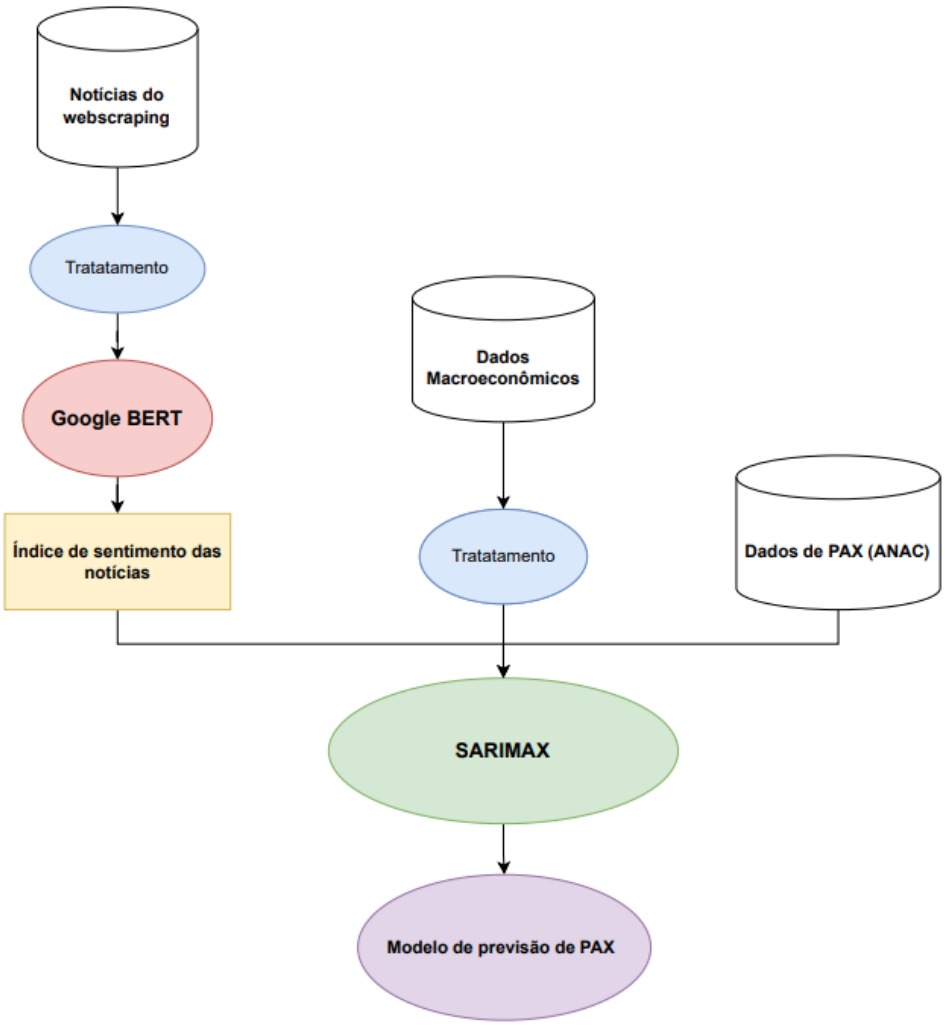


FIGURA 3.1 – Diagrama da metodologia simplificada. Fonte: Autor

3.2 Web Scraping

De uma maneira geral, para o web scraping realizado foram seguidos os seguintes passos.

1. Identificar o site-alvo: Determine qual site de notícias serão coletadas as notícias e seus textos.
2. Analisar a estrutura do site: Foi inspecionado código-fonte do site para entender a estrutura HTML do site e poder executar comandos como apertar botões para mudar o mês e filtrar notícias por palavras específicas.
3. Escolher uma biblioteca ou ferramenta de web scraping: O web scraping foi realizado com python utilizando a biblioteca selenium, então foi executado o código.
4. Armazenar e processar os dados: As notícias foram extraídas como uma tabela para cada mês de cada ano. Em cada uma dessas tabelas existem três colunas, uma de data, uma de título da notícia e o url da notícia.

Um aspecto de suma relevância para a execução do web scraping e a concepção do presente estudo refere-se à determinação criteriosa das palavras-chave utilizadas como filtros para a seleção das notícias. A intenção subjacente consiste em identificar reportagens que descrevam cenários econômicos tanto em âmbito nacional quanto internacional, bem como contextos relacionados ao turismo e outros fatores capazes de influenciar e moldar a demanda por passagens aéreas.

Essa abordagem permite uma análise aprofundada e abrangente do panorama econômico, possibilitando a compreensão dos elementos que permeiam o setor do turismo e os diversos fatores que exercem influência sobre a procura por serviços de transporte aéreo. Dessa forma, ao delimitar criteriosamente as palavras-chave empregadas como critérios de seleção, torna-se possível obter um conjunto de notícias que reflete de maneira representativa a conjuntura econômica global, regional e setorial, proporcionando uma base sólida para a análise subsequentemente realizada.

3.3 Tratamento dos dados de notícias

Após o web scraping, os links das notícias são salvos, em seguida passam por um tratamento de dados para serem armazenados em uma tabela com todos os links de todas as palavras chave com suas respectivas datas e links tratados.

Em seguida, a partir dos links de notícias armazenados utilizando o web scraping, captura-se então os textos das. Portanto, acrescenta-se a coluna de texto à tabela de

In [6]: `links_final`

Out[6]:

	Data	Titulo	Link	LinkLimpo	Data_Scraping	Tag
0	2023-05-08	\n Entenda como a nova regra de reajust...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia
1	2023-05-08	\n O Assunto #955: A política de valori...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia
2	2023-05-01	\n 'Faz mais de um ano que trabalho do ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia
3	2023-05-03	\n Haddad pede a Lira votação do arcabo...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia
4	2023-05-04	\n Copom: o que o BC ainda espera para ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia
...	...	...	...	...	...	...
3	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200606	turismo
0	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo
1	2023-06-02	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo
2	2022-09-13	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo
3	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo

8854 rows × 6 columns

FIGURA 3.2 – Tabela dos links de notícias e dados respectivos. Fonte: Autor

dados anteriormente apresentada e essa tabela final de dados brutos é armazenada em um arquivo de formato csv, com o intuito de reter uma base de dados bruta sem pré tratamentos e que passou apenas por coleta de dados para trabalhos e tratamentos futuros.

In [8]: `links_final`

Out[8]:

	Data	Titulo	Link	LinkLimpo	Data_Scraping	Tag	texto
0	2023-05-08	\n Entenda como a nova regra de reajust...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	Agora em seu terceiro man...
1	2023-05-08	\n O Assunto #955: A política de valori...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	Você pode ouvir O Assunto...
2	2023-05-01	\n 'Faz mais de um ano que trabalho do ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rFuncionários estão deixando...
3	2023-05-03	\n Haddad pede a Lira votação do arcabo...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rO ministro da Fazenda, Fern...
4	2023-05-04	\n Copom: o que o BC ainda espera para ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rPresidente do Banco Central...
...	...	...	...	...	...	...	...
3	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200606	turismo	Frango ao molho pardo, tutu de fe...
0	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo	Frango ao molho pardo, tutu de fe...
1	2023-06-02	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo	As belezas naturais de Ponta Gros...
2	2022-09-13	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo	1 de 4\rA capital conta com dezenas...
3	2022-04-25	NaN	https://g1.globo.com/busca/click?q=turismo&p=1...	https://g1.globo.com/busca/click?q=turismo&p=1...	200605	turismo	"O que mais atrai o turista para ...

8854 rows × 7 columns

FIGURA 3.3 – Tabela final de dados brutos. Fonte: Autor

Ainda prosseguindo com o tratamento, faz-se a limpeza do texto para prepará-lo para o modelo do Google BERT. Essa limpeza trata-se da da retirada de palavras que não irão contribuir com o resultado do modelo de linguagem natural, além disso, também é feita uma padronização do texto em ASCII.

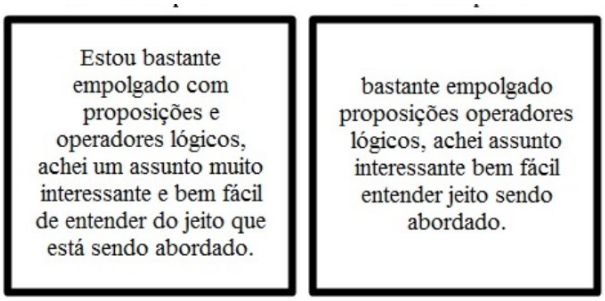


FIGURA 3.4 – Textos antes e depois de remoção de palavras. Fonte: Autor

	Data	Titulo	Link	LinkLimpo	Data_Scraping	Tag	texto	texto_limpo
0	2023-05-08	\n Entenda como a nova regra de reajust...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	Agora em seu terceiro man...	agor terceir mandato president luiz inaci lul ...
1	2023-05-08	\n O Assunto #955: A política de valori...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	Você pode ouvir O Assunto...	pod ouv assunt g1 globoplay spotify castbox go...
2	2023-05-01	\n 'Faz mais de um ano que trabalho do ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rFuncionários estão deixando...	1 1 funcionari deix inform empres viaj mud def...
3	2023-05-03	\n Haddad pede a Lira votação do arcabo...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rO ministro da Fazenda, Fern...	1 1 ministr fazenda fern haddad imag 10 abril ...
4	2023-05-04	\n Copom: o que o BC ainda espera para ...	https://g1.globo.com/busca/click?q=macroeconom...	https://g1.globo.com/busca/click?q=macroeconom...	202305	macroeconomia	1 de 1\rPresidente do Banco Central...	1 1 president banc central camp neto ministr f...

FIGURA 3.5 – Tabela com a coluna de texto processado. Fonte: Autor

3.4 Análise de sentimento com o Google BERT

3.4.1 Introdução ao modelo

O modelo do Google BERT (Bidirectional Encoder Representations from Transformers) foi feito em 2018 e utiliza redes neurais com arquitetura conhecidas como Transformer, de forma eficiente e capturar relações de dependência de longo alcance.

A vantagem em se utilizar o modelo BERT é o fato de ser um modelo pré-treinado, o que significa que foi treinado em grandes quantidades de texto não rotulado de diversas fontes, como livros, artigos e sites. Enquanto o treinamento ocorre, o modelo otimiza a captura de informações contextuais e semânticas entre as palavras e as relações delas em uma sentença.

Uma das principais características do BERT é a sua capacidade de processar texto em uma direção bidirecional. Isso significa que ele leva em consideração o contexto das palavras tanto à esquerda quanto à direita de cada palavra em uma frase. Essa abordagem bidirecional permite que o modelo capture nuances e dependências contextuais mais complexas.

O treinamento do BERT é baseado em duas tarefas principais: pré-treinamento e

ajuste fino (fine-tuning). No pré-treinamento, o modelo é treinado em tarefas como previsão de palavras mascaradas (Masked Language Model - MLM) e previsão de próxima sentença (Next Sentence Prediction - NSP). Essas tarefas ajudam o modelo a aprender a representação contextual das palavras e a entender a relação entre as sentenças.

Após o pré-treinamento, o BERT passa por um processo de ajuste fino em tarefas específicas de PLN, como classificação de sentimentos, identificação de entidades nomeadas e resposta a perguntas. Durante o ajuste fino, o modelo é treinado em um conjunto de dados rotulados e adaptado para a tarefa específica em questão.

3.4.2 Tokenização de textos

A tokenização é fundamental e amplamente utilizada em modelos de processamento de linguagem natural. Esse processo envolve a divisão do texto em unidades menores chamadas ”tokens”, os quais podem ser palavras individuais, partes de palavras ou até mesmo caracteres.

Esse processo é essencial porque o BERT processa informações em sequências de tokens, considerando o contexto de cada token em relação aos tokens anteriores e posteriores. Portanto, a tokenização é o processo de dividir um texto em partes menores, permitindo que o modelo capture as relações contextuais entre as palavras.

O BERT utiliza uma técnica chamada WordPiece tokenization, na qual palavras são divididas em subtokens, que podem ser fragmentos de palavras ou palavras completas. Por exemplo, a palavra ”desigualdade”pode ser dividida em ”desi”, ”##gua”, ”##lda”e ”##des”. Os subtokens que não representam o início de uma palavra são prefixados com ”##”para indicar que eles pertencem a uma palavra maior. Esse processo permite que o BERT lide com um vocabulário amplo e capture a estrutura interna das palavras.

Importância da Tokenização no Google BERT:

1. **Compreensão de Contexto:** A tokenização permite que o BERT considere o contexto mais amplo de uma sentença, melhorando a compreensão das relações entre palavras.

2. **Manejo de Vocabulário:** Dividir palavras em subtokens ajuda o BERT a lidar com um conjunto limitado de tokens no vocabulário, incluindo palavras raras e complexas.

3. **Codificação Posicional:** Atribuir posições únicas a cada token possibilita que o BERT entenda a ordem das palavras na sequência.

4. **Captura de Composição de Palavras:** A tokenização de subtokens permite que o BERT compreenda a formação de novos termos a partir de partes de palavras.

5. **Generalização e Transferência de Aprendizado:** A tokenização contribui para a capacidade do BERT de generalizar e aplicar conhecimento a diversas tarefas de PLN.

Texto tratado	Tokens
dolar fechou leve alta nesta quintafeira 4 abrir pregao baixa dia apos decisao comite politica monetaria copom manter selic taxa basica juros 1375 ano ontem tambem federal reserve fed banco central americano elevou juros 025...	dolar, fe, ##cho, ##u, leve, alta, nesta, quinta, ##fe, ##ira, 4, abrir, pre, ##gao, baixa, dia, apos, decisao, comite, politica, moneta, ##ria, cop, ##om, manter, sel, ##ic, taxa, basica, juros, 1375, ano, ont, ##em, tambem, federal, reserve, fed, banco, central, americano, elev, ##ou, juros, 025,...

FIGURA 3.6 – Exemplo de texto e seus tokens gerados pelo modelo BERT. Fonte: Autor

3.4.3 Embedding de palavras

O processo de embedding desempenha uma função crucial dentro do arcabouço do modelo Google BERT (Bidirectional Encoder Representations from Transformers). Este modelo utiliza a técnica de embedding para representar as unidades textuais, como palavras ou subtokens, em um espaço vetorial de alta dimensionalidade. Essa etapa é fundamental para a captura de informações semânticas e contextuais das palavras presentes no corpus de treinamento.

A operação de embedding no BERT ocorre durante a fase de pré-processamento dos dados, na qual os textos são segmentados em subtokens e, posteriormente, associados a vetores numéricos. Essa conversão é efetuada por meio de uma camada de embedding, que converte as representações discretas das palavras em representações contínuas no espaço vetorial. Essas representações são aprimoradas simultaneamente ao treinamento global do modelo BERT durante a etapa de pré-treinamento.

No contexto do BERT, a técnica de embedding é bidirecional, o que significa que ela considera tanto os elementos que precedem quanto os que sucedem um determinado subtoken. Esse enfoque permite que o modelo apreenda relações contextuais complexas, ampliando a riqueza da representação vetorial de cada subtoken. Adicionalmente, o BERT incorpora mecanismos de posicionamento, nos quais cada subtoken é associado a uma posição única no texto, facilitando a manutenção da estrutura e ordem das palavras.

Em resumo, a operação de embedding no Google BERT é um estágio fundamental na geração de representações semânticas avançadas para palavras e subtokens. Essas representações são de suma importância para a eficácia do modelo em diversas tarefas

relacionadas ao Processamento de Linguagem Natural, permitindo uma compreensão profunda e eficaz do contexto e semântica presentes nos textos.

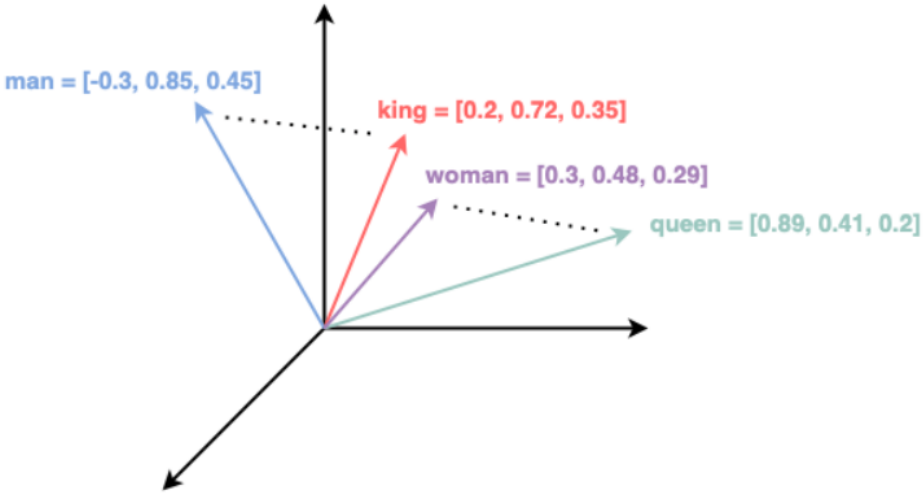


FIGURA 3.7 – Exemplo de embeddings com três dimensões. Fonte: (BAELDUNG, 2023)

Seguindo essa ideia, o resultado obtido é um vetor para cada token extraído do texto tratado. No entando, é aplicada uma abordagem de calcular a média, em cada dimensão, dos vetores de tokens e isso emerge como um procedimento substancial na busca por uma representação concisa e agregada da semântica inerente aos textos. Através dessa metodologia, as representações semânticas individuais atribuídas a cada token são amalgamadas em um único vetor composto. A presente técnica visa oferecer uma síntese que encapsula as nuances semânticas predominantes do texto, permitindo uma visão panorâmica das informações subjacentes.

Dentro desse contexto, a média dos vetores de tokens atua como um mecanismo de redução da dimensionalidade do espaço vetorial, buscando capturar as características semânticas mais relevantes do texto e tende a atenuar detalhes específicos, direcionando o foco para os elementos centrais que compõem a essência semântica do texto.

No entanto, na aplicação dessa abordagem, há uma consideração de custo-benefício entre a simplificação da representação e a preservação da riqueza semântica. A eliminação de detalhes menores resulta na perda de informações contextuais mais intricadas, portanto, essa aplicação torna-se aplicável ao passo que o objetivo está na apreensão das características gerais do texto, em detrimento das particularidades intrínsecas de cada palavra. Em suma, a abordagem presente apresentará um viés em direção à síntese da semântica textual, explorando a balança entre representação simplificada e a captura dos aspectos predominantes da narrativa.

O resultado obtido pela média desses valores foi, então, normalizado de -1 a 1 , de modo a se criar uma variável que foi chamada de $variavel_bert_1$, enquanto ao dividirmos

essa média pela quantidade de notícias presentes em cada mês do ano para se ponderar a intensidade da variável pela quantidade de texto analisado, obteve-se a $variavel_bert_2$. Em seguida, a série histórica desse valor direcional da semântica do texto foi analisada a partir de gráficos.

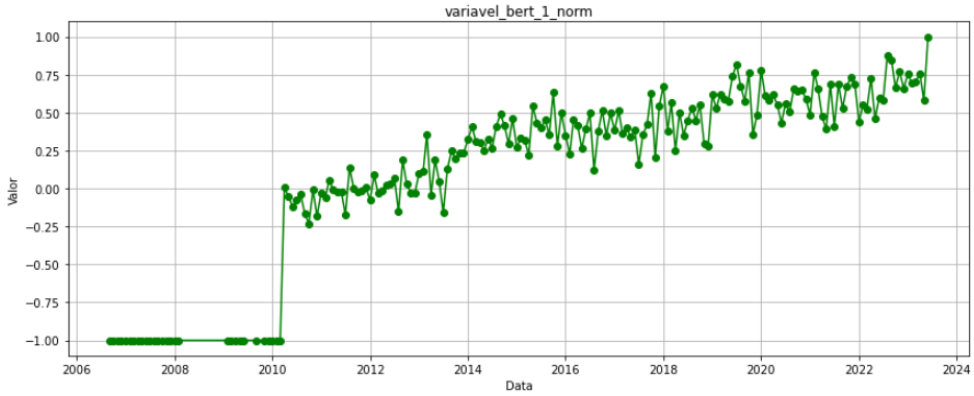


FIGURA 3.8 – Resultado final da média das dimensões dos vetores. Fonte: Autor

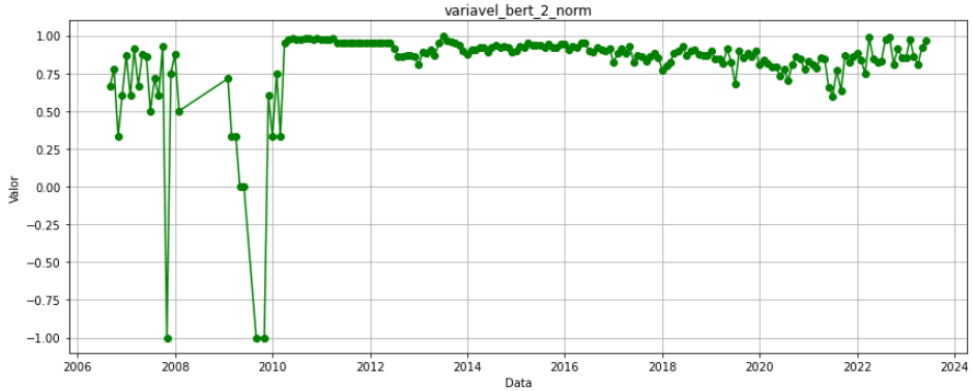


FIGURA 3.9 – Resultado final da média das dimensões dos vetores ponderada pela quantidade de notícias. Fonte: Autor

Consequentemente, a interpretação dos gráficos sugere a presença de anomalias nos dados precedentes ao ano de 2012. Esta irregularidade nos registros pode ser atribuída, com alta probabilidade, à frequência reduzida de publicações pelo site de notícias em períodos anteriores, resultando em uma coleta de dados menos robusta. Este fator limitante influencia diretamente a integridade e a confiabilidade dos resultados analíticos obtidos pelo modelo. Adicionalmente, a evolução temporal das estratégias de comunicação e a mudança nos padrões de publicação podem ter contribuído significativamente para as discrepâncias observadas. Portanto, será considerado o período a partir do ano de 2012 para análises subsequentes.

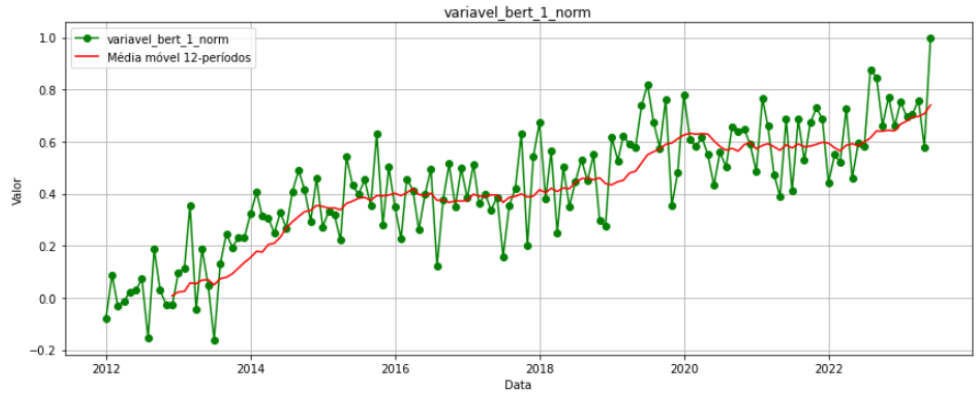


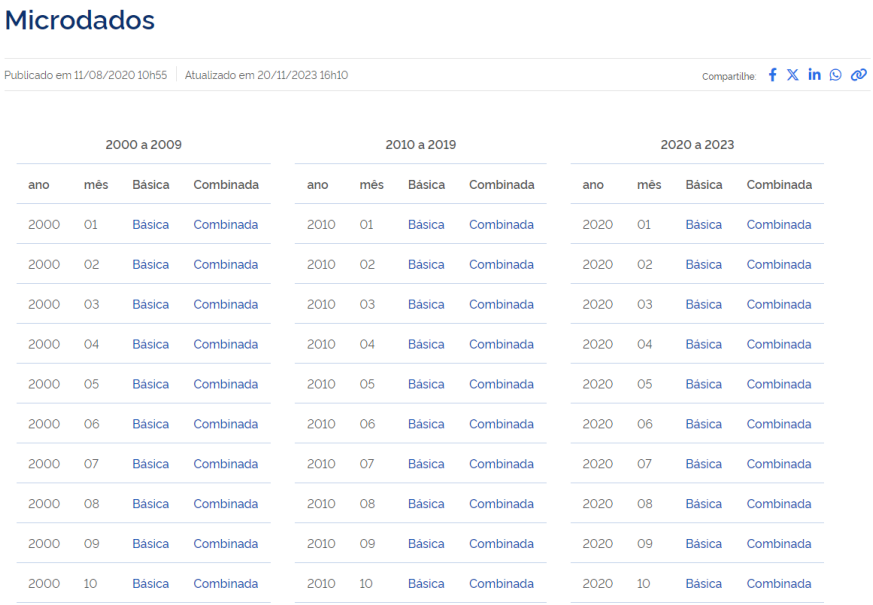
FIGURA 3.10 – Resultado final da média das dimensões dos vetores considerando a partir de 2012. Fonte: Autor

FIGURA 3.11 – Resultado final da média das dimensões dos vetores ponderada pela quantidade de notícias considerando a partir de 2012. Fonte: Autor

3.5 Dados de demanda de passageiros

Os dados referentes à demanda de passageiros foram coletados a partir das tabelas de dados anuais fornecidas pela Agência Nacional de Aviação Civil (ANAC). Este processo envolveu uma análise detalhada e a organização das informações disponibilizadas. Para garantir a precisão e a relevância dos dados, realizou-se um tratamento cuidadoso das tabelas. Essa etapa foi essencial para consolidar os números de forma que representassem a totalidade de passageiros relacionados a uma determinada data. A soma cumulativa dos passageiros, obtida através deste método, oferece uma visão abrangente e fiel das tendências e variações na demanda de passageiros ao longo do período analisado. Esse conjunto de dados tratados é fundamental para compreender as dinâmicas do setor aéreo e auxilia na tomada de decisões baseadas em informações concretas e bem fundamentadas.

Após a conclusão do processo de tratamento dos dados, e focalizando a análise nos intervalos temporais subsequentes ao ano de 2012, identifica-se uma tendência incremental moderada na demanda de passageiros aéreos entre 2012 e 2015. Entretanto, o ano de 2016 é marcado por uma diminuição nessa demanda, correlacionada ao intensificar de uma crise econômica no contexto brasileiro. Seguiu-se um período de recuperação e cres-



Microdados

Publicado em 11/08/2020 10h55 | Atualizado em 20/11/2023 16h10 Compartilhe:

2000 a 2009

ano	mês	Básica	Combinada
2000	01	Básica	Combinada
2000	02	Básica	Combinada
2000	03	Básica	Combinada
2000	04	Básica	Combinada
2000	05	Básica	Combinada
2000	06	Básica	Combinada
2000	07	Básica	Combinada
2000	08	Básica	Combinada
2000	09	Básica	Combinada
2000	10	Básica	Combinada

2010 a 2019

ano	mês	Básica	Combinada
2010	01	Básica	Combinada
2010	02	Básica	Combinada
2010	03	Básica	Combinada
2010	04	Básica	Combinada
2010	05	Básica	Combinada
2010	06	Básica	Combinada
2010	07	Básica	Combinada
2010	08	Básica	Combinada
2010	09	Básica	Combinada
2010	10	Básica	Combinada

2020 a 2023

ano	mês	Básica	Combinada
2020	01	Básica	Combinada
2020	02	Básica	Combinada
2020	03	Básica	Combinada
2020	04	Básica	Combinada
2020	05	Básica	Combinada
2020	06	Básica	Combinada
2020	07	Básica	Combinada
2020	08	Básica	Combinada
2020	09	Básica	Combinada
2020	10	Básica	Combinada

FIGURA 3.12 – Seção do site da ANAC para coleta de dados. Fonte: Autor

cimento no número de passageiros, persistindo até a emergência da crise sanitária global provocada pela COVID-19. Esta última ocasionou uma redução substancial na demanda por transporte aéreo, desencadeando um período de recuperação que ainda está em curso.

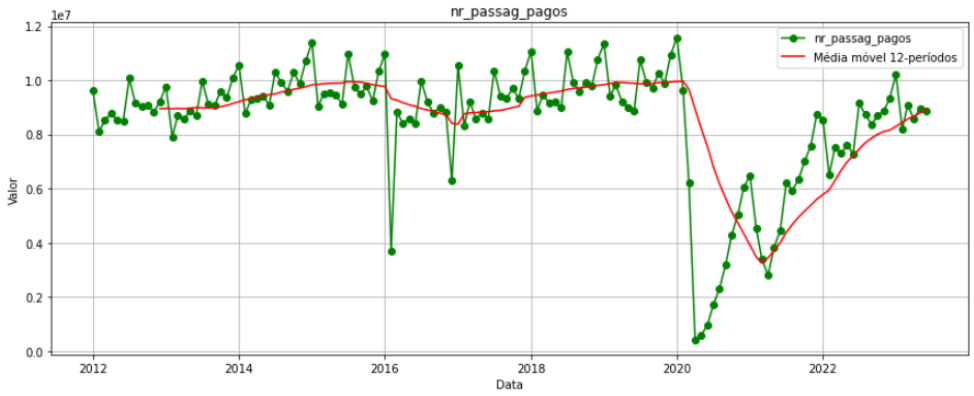


FIGURA 3.13 – Série histórica de número de passageiros pagantes de vôos. Fonte: Autor

3.6 Variáveis macroeconômicas

3.6.1 PIB

Os dados referentes ao Produto Interno Bruto (PIB) foram adquiridos de forma mensal, diretamente do Banco Central, através do portal Ipeadata. Este repositório é amplamente reconhecido por disponibilizar informações econômicas oficiais e atualizadas, essenciais para análises econômicas detalhadas e confiáveis. A utilização de dados provenientes de

uma fonte oficial e respeitada, como o Banco Central, assegura a precisão e a relevância das informações econômicas empregadas no estudo.

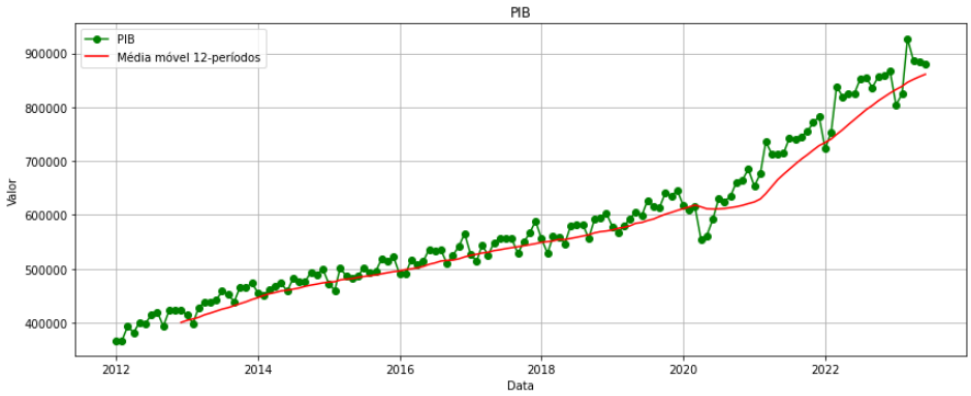


FIGURA 3.14 – Série histórica do PIB mensal. Fonte: Autor

3.6.2 Índice de confiança do consumidor

Os dados referentes ao Índice de Confiança do Consumidor foram coletados em uma base mensal, provenientes do Banco Central do Brasil. Estes dados foram acessados através do portal de Dados Abertos do Banco Central, especificamente na página Índice de Confiança do Consumidor. A escolha desta fonte assegura a confiabilidade e a validade dos dados utilizados na análise, dada a credibilidade e a autoridade do Banco Central como instituição fornecedora de informações econômicas fundamentais e atualizadas para pesquisas no âmbito econômico.

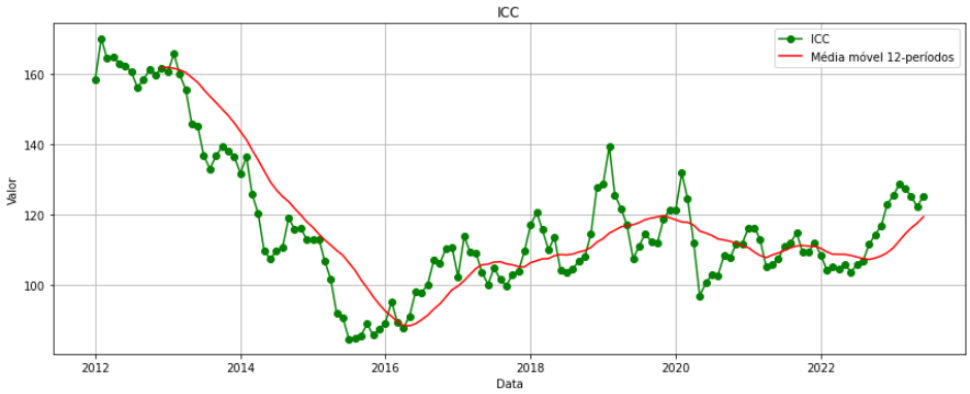


FIGURA 3.15 – Série histórica do índice de confiança do consumidor mensal. Fonte: Autor

4 Resultados e discussões

4.1 Análise da série temporal de demanda por passagem aérea

4.1.1 Decomposição da série temporal

Com o intuito de se estudar a série histórica de número de passageiros pagos, foi feita uma decomposição da série temporal, uma técnica estatística utilizada para analisar as variações intrínsecas dentro dos dados ao longo do tempo.

A série original, posicionada no gráfico superior, compreende os dados integrais que registram o fluxo total de passageiros pagos ao longo do intervalo de tempo estudado.

O segundo gráfico, intitulado ”Tendência”, expõe a evolução subjacente da série ao longo do tempo. Este componente é calculado para determinar a trajetória de longo prazo do número de passageiros, abstraindo as variações sazonais e outras oscilações de curta duração, e é crucial para entender as direções estratégicas do tráfego aéreo.

A componente ”Sazonalidade”identifica padrões periódicos específicos ao número de passageiros pagos, destacando a periodicidade com a qual as variações ocorrem dentro de um ciclo anual. Tais padrões podem ser influenciados por fatores como períodos de férias, eventos sazonais e preferências de viagem.

Por fim, a componente ”Resíduos”reflete a parcela da série que não é explicada pela tendência ou sazonalidade. Este resíduo pode ser atribuído a variações aleatórias, eventos esporádicos ou outras irregularidades que impactam o número de passageiros pagos e que não são capturadas pelos componentes sistemáticos.

Diante da análise da decomposição mencionada, é justificável incorporar um componente sazonal nos modelos a serem submetidos a testes e avaliações, uma vez que a presença de um padrão sazonal parece substancialmente evidente a partir dos dados observados. Essa inclusão permitirá uma modelagem mais precisa e abrangente das flutuações temporais subjacentes ao fenômeno em questão.

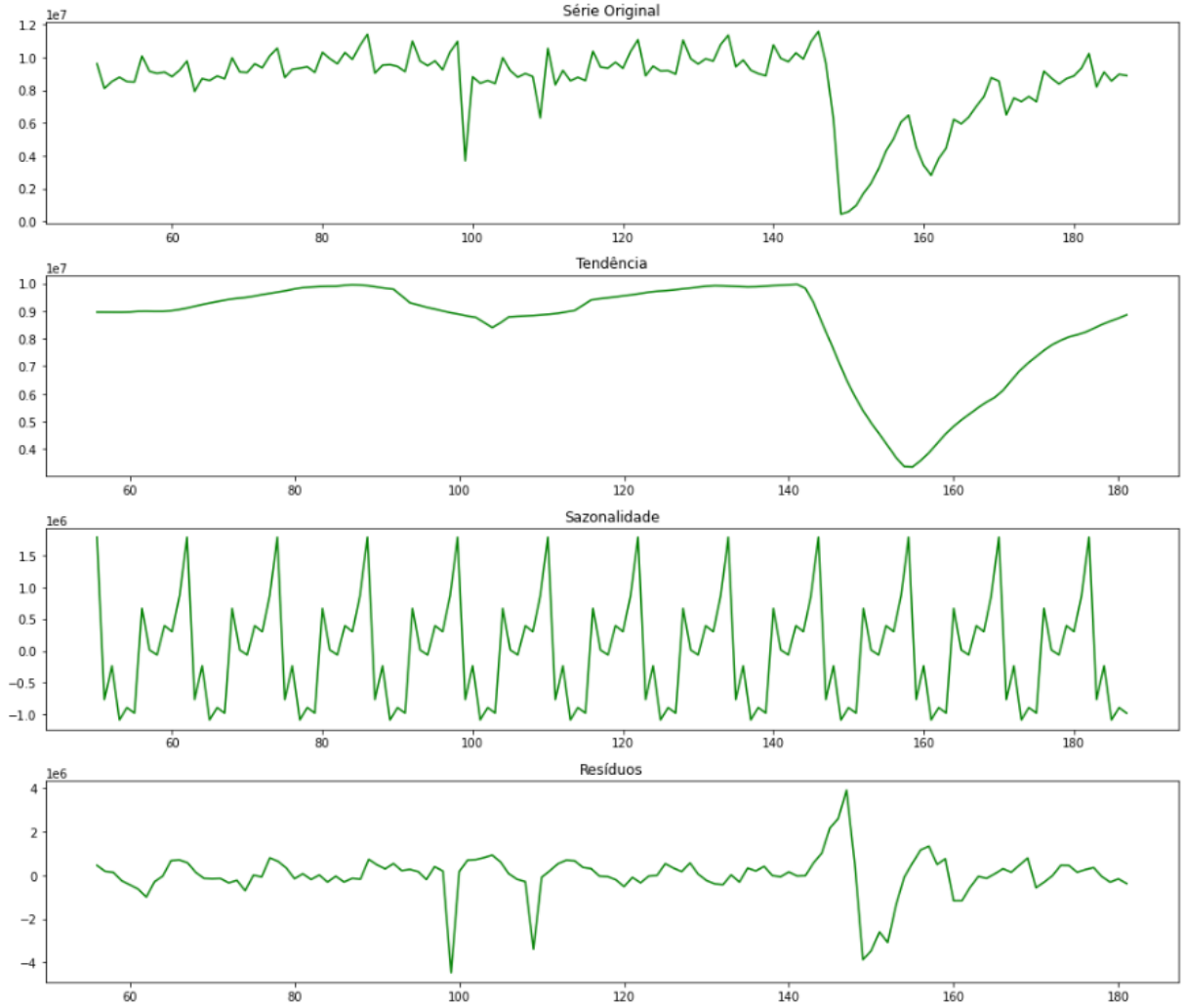


FIGURA 4.1 – Decomposição da série temporal. Fonte: Autor

4.1.2 Estacionariedade da série

O teste ADF (Augmented Dickey-Fuller) é um teste estatístico utilizado para avaliar a estacionariedade em séries temporais. A hipótese nula no teste ADF assume que a série temporal possui uma raiz unitária, o que implica que a série é não estacionária. Em outras palavras, a hipótese nula postula que a série exibe tendência e não possui estacionariedade.

Os níveis de significância no teste ADF representam o limiar de probabilidade que determina quando rejeitamos a hipótese nula. Normalmente, um nível de significância de 0,05 (ou 5%) é utilizado, o que significa que rejeitamos a hipótese nula somente se a probabilidade de obter os resultados observados for menor que 5% de chance sob a hipótese nula.

Não podemos descartar a hipótese nula no teste ADF quando a estatística de teste calculada for maior que os valores críticos correspondentes ao nível de significância escolhido. Isso indica que não há evidência estatística suficiente para rejeitar a hipótese nula, sugerindo que a série temporal em análise possui raiz unitária, ou seja, não é esta-

cionária. Portanto, o teste ADF ajuda a determinar se a série temporal é estacionária ou não com base na significância estatística da estatística de teste em relação aos níveis de significância predefinidos.

Valor Estatístico ADF	-2.572
Valor p	0.099
Valores Críticos (1%)	-3.484
Valores Críticos (5%)	-2.885
Valores Críticos (10%)	-2.579

1. Valor Estatístico ADF (-2.572): O valor estatístico ADF é menor que os valores críticos em 1%, 5% e 10%, o que sugere que a série pode ser estacionária. No entanto, esse valor por si só não é suficiente para concluir com confiança a estacionariedade.

2. Valor p (0.099): O valor p é maior que o nível de significância comum de 0.05. Isso indica que não temos evidências suficientes para rejeitar a hipótese nula de que a série não é estacionária. O valor p mais alto sugere que a série pode não ser estacionária.

3. Valores Críticos: Os valores críticos são usados para determinar se o valor estatístico ADF é significativamente menor que o esperado ao acaso. Neste caso, o valor estatístico ADF está próximo dos valores críticos, o que sugere que a série não é claramente estacionária.

Em resumo, com base nos resultados do teste ADF e no valor p, não podemos afirmar com confiança que a série temporal é estacionária. Os resultados indicam uma falta de evidências claras de estacionariedade, pois o valor p é maior que 0.05 e o valor estatístico ADF está próximo dos valores críticos. É possível que a série seja não estacionária ou que seja necessário realizar análises adicionais ou considerar outras informações para determinar sua natureza.

4.1.3 Transformações da série temporal

4.1.3.1 Logarítimica

A transformação logarítmica de uma série temporal envolve a aplicação do logaritmo a cada ponto de dados na série. Isso é feito para estabilizar a variância e reduzir o impacto de valores extremos, tornando a série mais adequada para análises estatísticas.

Valor Estatístico ADF	-3.709
Valor p	0.004
Valores Críticos (1%)	-3.479
Valores Críticos (5%)	-2.883
Valores Críticos (10%)	-2.578

1. Valor Estatístico ADF (-3.709): O valor estatístico ADF é menor que os valores críticos em 1%, 5% e 10%, indicando que a série parece ser estacionária.

2. Valor p (0.004): O valor p é menor que o nível de significância comum de 0.05, o que sugere que temos evidências estatisticamente significativas para rejeitar a hipótese nula de não estacionariedade.

3. Valores Críticos: Os valores críticos são usados para determinar se o valor estatístico ADF é significativamente menor que o esperado ao acaso. Neste caso, o valor estatístico ADF está abaixo dos valores críticos, indicando estacionariedade.

Em resumo, com base nos resultados do novo teste ADF, podemos concluir com mais confiança que a série temporal é estacionária, pois o valor estatístico ADF é menor que os valores críticos e o valor p é menor que 0.05.

4.1.4 Autocorrelações da série temporal

Para avaliar a autocorrelação na série temporal já feita a transformação logarítimca os gráficos de Autocorrelação (ACF) e Autocorrelação Parcial (PACF) tornam-se ferramentas relevantes, dada a possibilidade de identificar padrões de dependência temporal.

O gráfico ACF mede a correlação entre observações de uma série temporal e suas versões anteriores (lags). Se os valores do ACF são significativamente diferentes de zero (fora das bandas de confiança sombreadas), isso indica autocorrelação e, portanto, uma relação serial nos dados.

O gráfico PACF, por outro lado, mede a autocorrelação entre as observações e suas versões anteriores, controlando pelos valores intermediários. Valores significativos no PACF indicam a ordem potencial de um modelo autorregressivo (AR).

Na análise da série temporal em questão, a função de autocorrelação (ACF) inicialmente apresenta uma forte correlação positiva, que diminui rapidamente e se torna estatisticamente insignificante após o primeiro atraso (lag). Isso indica a presença de uma dependência significativa do valor anterior na série, embora essa dependência não seja sustentada ao longo de períodos prolongados.

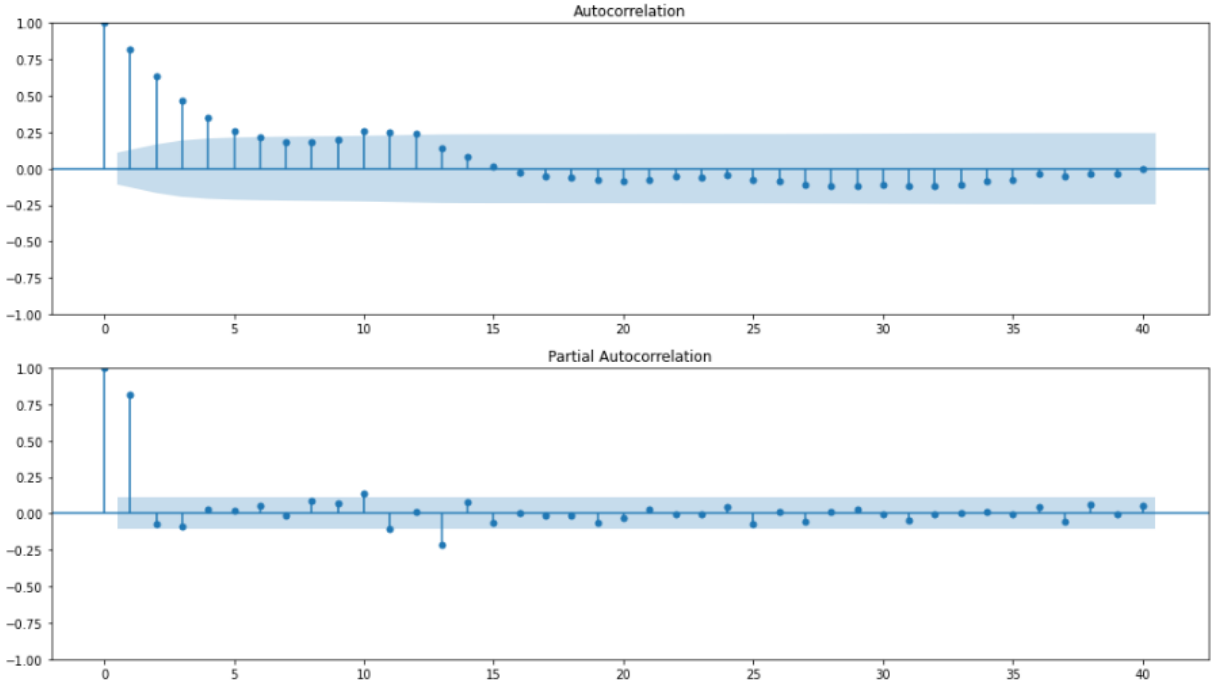


FIGURA 4.2 – Gráficos de autocorrelação. Fonte: Autor

Ao examinar a função de autocorrelação parcial (PACF), observamos um pico significativo no primeiro atraso. Esse resultado sugere a possibilidade de um modelo autoregressivo AR(1), no qual o valor subsequente na série é principalmente influenciado pelo valor imediatamente anterior.

4.2 Modelos de demanda utilizando o PIB como variável exógena e considerando a sazonalidade

4.2.1 Primeiro modelo

4.2.1.1 Representação matemática

$$\ln(\text{nr\_passag\_pagos}_t) = \alpha \ln(\text{nr\_passag\_pagos}_{t-1}) + \beta \ln(\text{PIB}_t) + \theta\epsilon_{t-1} + \Phi \ln(\text{nr\_passag\_pagos}_{t-12}) + \epsilon_t \tag{4.1}$$

- $Y_t = \ln(\text{nr\_passag\_pagos}_t)$: Representa o logaritmo natural do número de passageiros pagantes no tempo t.
- α: Coeficiente do termo autoregressivo (AR) de primeira ordem. Reflete a relação de Y_t com seu valor no período anterior (Y_{t-1}).

- $X_t = \ln(\text{PIB}_t)$: Representa o logaritmo natural do Produto Interno Bruto no tempo t, atuando como variável exógena.
- β: Coeficiente da variável exógena X_t. Representa a influência do PIB no valor atual de Y.
- θ: Coeficiente do termo de média móvel (MA) de primeira ordem. Indica a relação entre o erro atual (ϵ_t) e o erro no período anterior (ϵ_{t-1}).
- Φ: Coeficiente do termo autoregressivo sazonal. Representa a influência do valor de Y no mesmo período do ano anterior (Y_{t-12}), considerando um ciclo sazonal de 12 períodos.
- ϵ_t: Termo de erro no tempo t, representando a parte do valor de Y_t que não é explicada pelo modelo.

4.2.1.2 Análise do modelo

SARIMAX Results

Dep. Variable:	nr_passag_pagos_log	No. Observations:	132
Model:	SARIMAX(1, 0, 1)x(1, 0, [], 12)	Log Likelihood	-17.493
Date:	Thu, 23 Nov 2023	AIC	44.987
Time:	03:20:58	BIC	59.401
Sample:	0	HQIC	50.844
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
PIB_log	1.2017	0.026	45.893	0.000	1.150	1.253
ar.L1	0.8462	0.147	5.743	0.000	0.557	1.135
ma.L1	0.0658	0.158	0.415	0.678	-0.245	0.376
ar.S.L12	0.2385	0.103	2.315	0.021	0.037	0.440
sigma2	0.0751	0.003	26.632	0.000	0.070	0.081

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	16896.95
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	20.72	Skew:	-6.13
Prob(H) (two-sided):	0.00	Kurtosis:	57.05

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURA 4.3 – Sumário do primeiro modelo. Fonte: Autor

Análise dos Testes Estatísticos:

- Os valores de z e os p-valores associados sugerem que os coeficientes PIB_log, ar.L1 e ar.S.L12 são estatisticamente significativos ao nível de 5%.

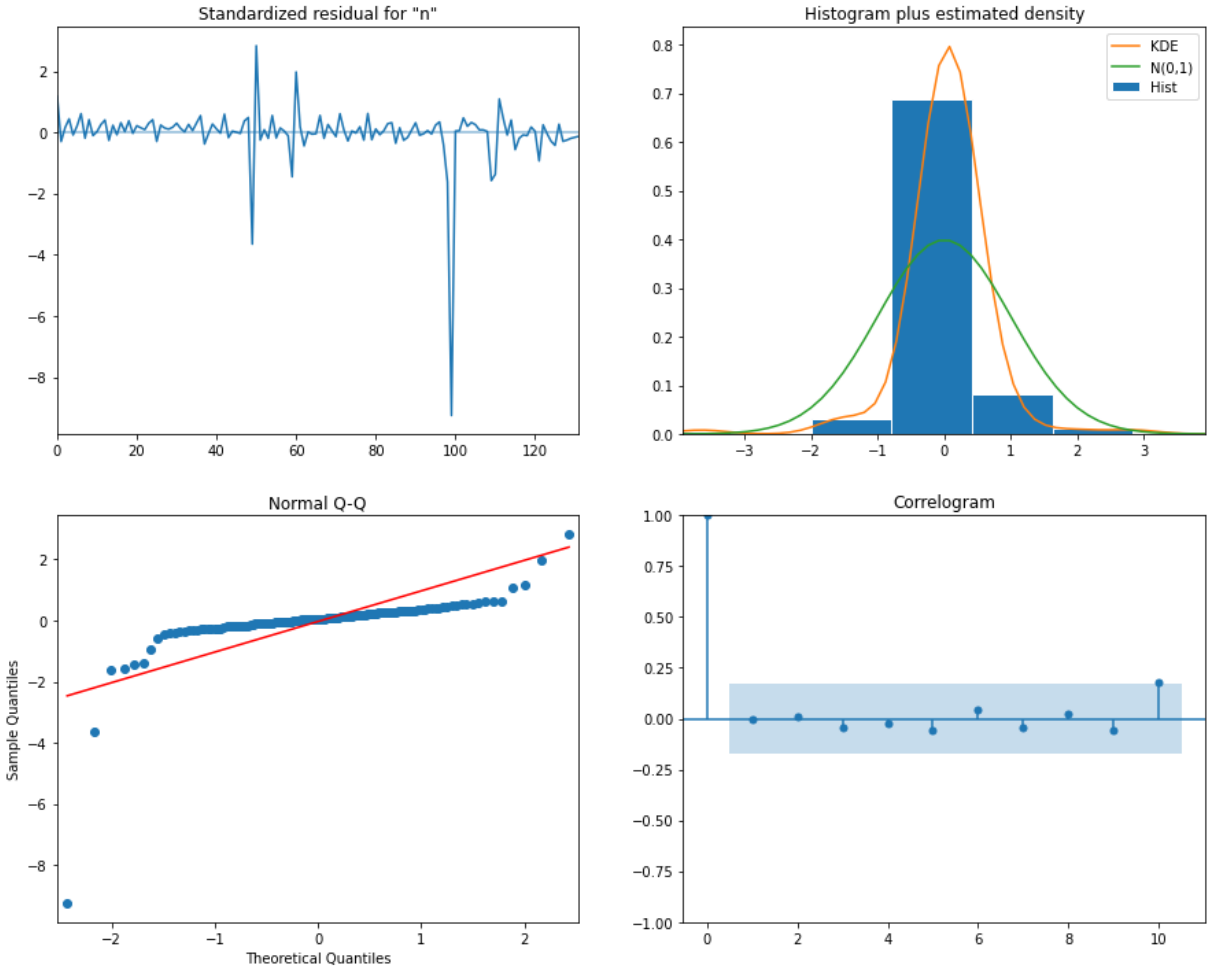


FIGURA 4.4 – Diagnóstico do primeiro modelo. Fonte: Autor

- O teste de Ljung-Box com um p-valor de 0.96 indica que não há autocorrelação significativa nos resíduos.
- O teste de Jarque-Bera apresenta um p-valor extremamente baixo, apontando para a não normalidade dos resíduos, o que é corroborado pela alta curtose e assimetria negativa.

Considerações Adicionais:

- O modelo parece capturar adequadamente a estrutura de dependência temporal dos dados.
- A não normalidade dos resíduos pode requerer investigações adicionais e potenciais ajustes no modelo.

4.2.2 Segundo modelo

4.2.2.1 Representação Matemática

O modelo SARIMAX especificado pode ser matematicamente representado pela seguinte equação:

$$\ln(\text{nr\_passag\_pagos}_t) = \alpha \ln(\text{nr\_passag\_pagos}_{t-1}) + \beta \ln(\text{PIB}_t) + \Phi \ln(\text{nr\_passag\_pagos}_{t-12}) + \epsilon_t \tag{4.2}$$

Onde:

- $\ln(\text{nr\_passag\_pagos}_t)$: Logaritmo natural do número de passageiros pagantes no tempo t.
- α: Coeficiente do termo autoregressivo de primeira ordem (AR(1)), indicando o efeito do valor logarítmico dos passageiros pagantes no período anterior.
- $\ln(\text{PIB}_t)$: Logaritmo natural do Produto Interno Bruto no tempo t, utilizado como variável exógena.
- β: Coeficiente estimado para a variável exógena $\ln(\text{PIB}_t)$, refletindo a influência do PIB sobre o número de passageiros pagantes.
- ϵ_t: Termo de erro estocástico no tempo t.

4.2.2.2 Análise do Modelo

As figuras a seguir apresentam o sumário alguns resultados do segundo modelo:

Análise dos Testes Estatísticos:

A partir dos resultados do modelo, podemos observar que:

- O coeficiente para $\ln(\text{PIB}_t)$ é estatisticamente significativo com um valor de z de 47.919 e um p-valor praticamente nulo, sugerindo uma forte influência do PIB sobre o número de passageiros pagantes.
- O termo autoregressivo α também é significativo, com um valor de z de 12.035 e um p-valor de 0.000.
- O valor do p-valor para o teste de Ljung-Box é 0.60, indicando que não há autocorrelação significativa nos resíduos do modelo.

SARIMAX Results

Dep. Variable:	nr_passag_pagos_log	No. Observations:	132
Model:	SARIMAX(1, 0, 0)x(1, 0, 0, 12)	Log Likelihood	-17.712
Date:	Thu, 23 Nov 2023	AIC	43.424
Time:	03:50:10	BIC	54.956
Sample:	0	HQIC	48.110
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
PIB_log	1.2019	0.025	47.919	0.000	1.153	1.251
ar.L1	0.8631	0.072	12.035	0.000	0.723	1.004
ar.S.L12	0.2333	0.101	2.319	0.020	0.036	0.430
sigma2	0.0754	0.003	27.008	0.000	0.070	0.081

Ljung-Box (L1) (Q):	0.28	Jarque-Bera (JB):	17798.65
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	22.25	Skew:	-6.30
Prob(H) (two-sided):	0.00	Kurtosis:	58.47

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURA 4.5 – Sumário do segundo modelo com apenas o PIB como variável exógena. Fonte: Autor

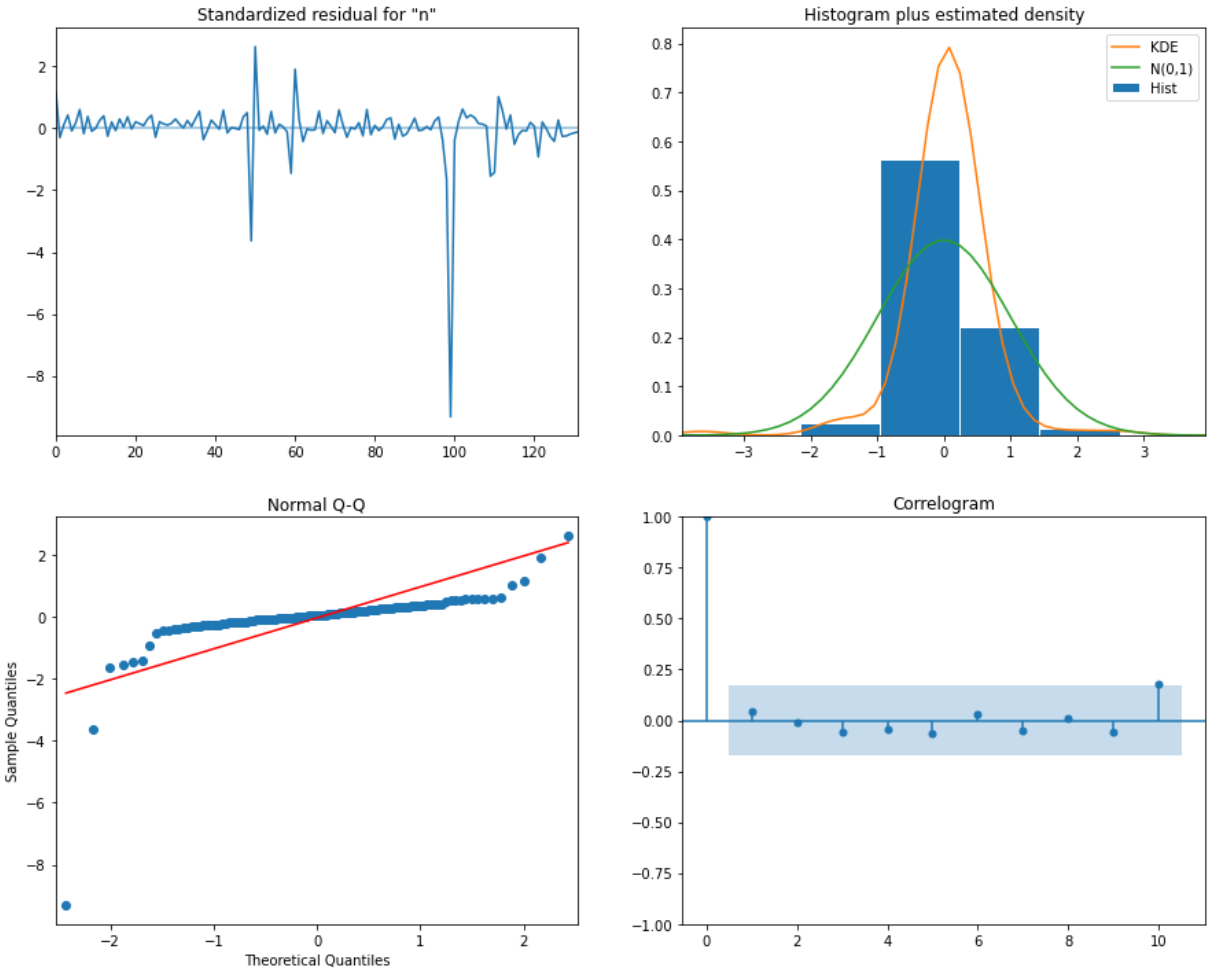


FIGURA 4.6 – Diagnóstico do segundo modelo com apenas o PIB como variável exógena. Fonte: Autor

```
MSE: 0.023647891259031436
RMSE: 0.1537787087311876
MAE: 0.14898792415964338
```

FIGURA 4.7 – Erros obtidos no segundo modelo com apenas o PIB como variável exógena. Fonte: Autor

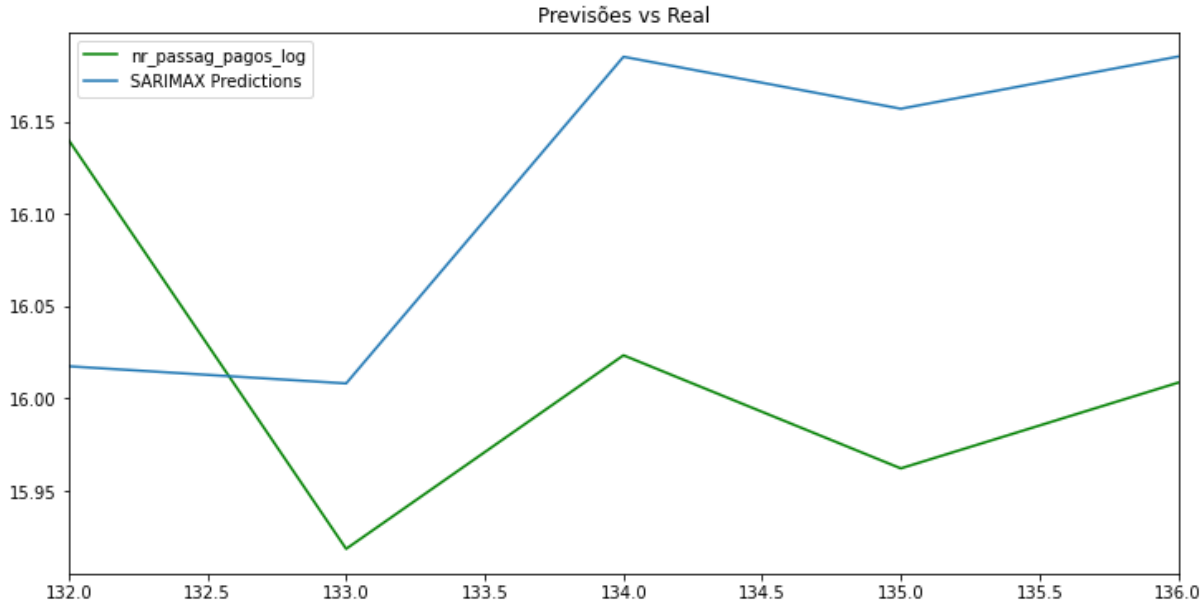


FIGURA 4.8 – Curvas de previsão do segundo modelo com apenas o PIB como variável exógena versus o resultado real. Fonte: Autor

- O teste de Jarque-Bera retorna um p-valor de 0.00, sugerindo que os resíduos não seguem uma distribuição normal.

Considerações Adicionais:

- A significância dos coeficientes indica que tanto os termos autoregressivos quanto as variáveis exógenas são relevantes para o modelo.
- A falta de autocorrelação nos resíduos sugere que o modelo está capturando adequadamente a dependência temporal dos dados.

4.2.3 Terceiro modelo

4.2.3.1 Representação Matemática

O modelo SARIMA aplicado pode ser representado pela seguinte equação:

$$\ln(\text{nr\_passag\_pagos}_t) = \alpha \ln(\text{nr\_passag\_pagos}_{t-1}) + \beta \ln(\text{PIB}_t) + \epsilon_t \tag{4.3}$$

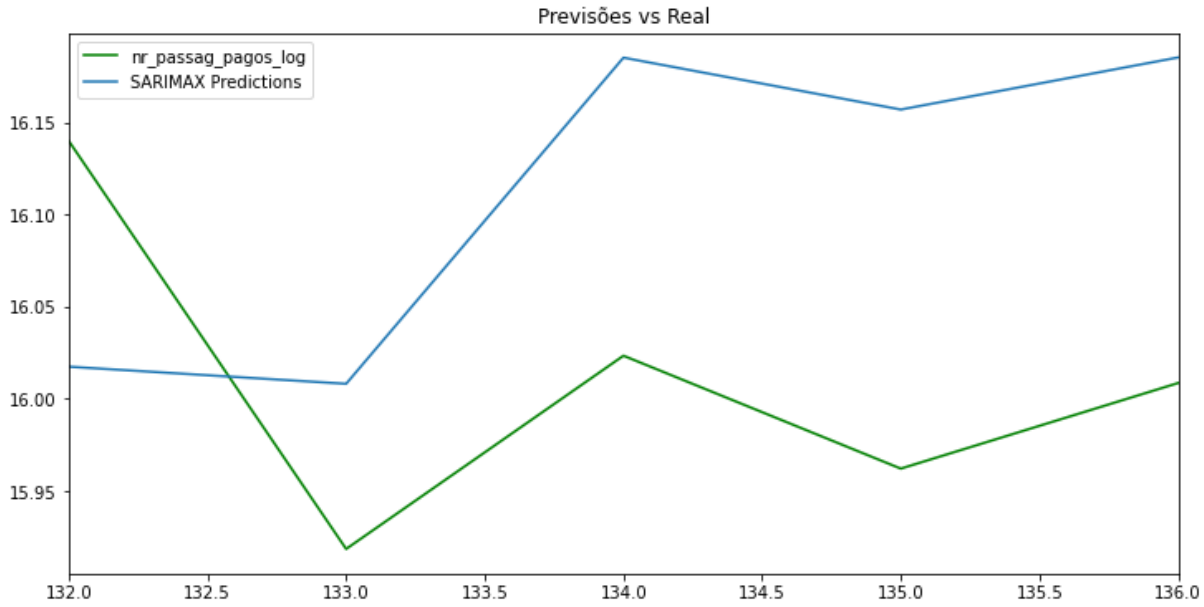


FIGURA 4.9 – Curvas de previsão do terceiro modelo com apenas o PIB como variável exógena versus o resultado real. Fonte: Autor

onde:

- $\ln(\text{nr\_passag\_pagos}_t)$: Logaritmo natural do número de passageiros pagantes no tempo t.
- α: Coeficiente do termo autoregressivo de primeira ordem, indicando o efeito do valor logarítmico dos passageiros pagantes no período anterior.
- $\ln(\text{PIB}_t)$: Logaritmo natural do Produto Interno Bruto no tempo t, utilizado como variável exógena.
- β: Coeficiente estimado para a variável exógena, refletindo a influência do PIB sobre o número de passageiros pagantes.
- ϵ_t: Termo de erro estocástico no tempo t.

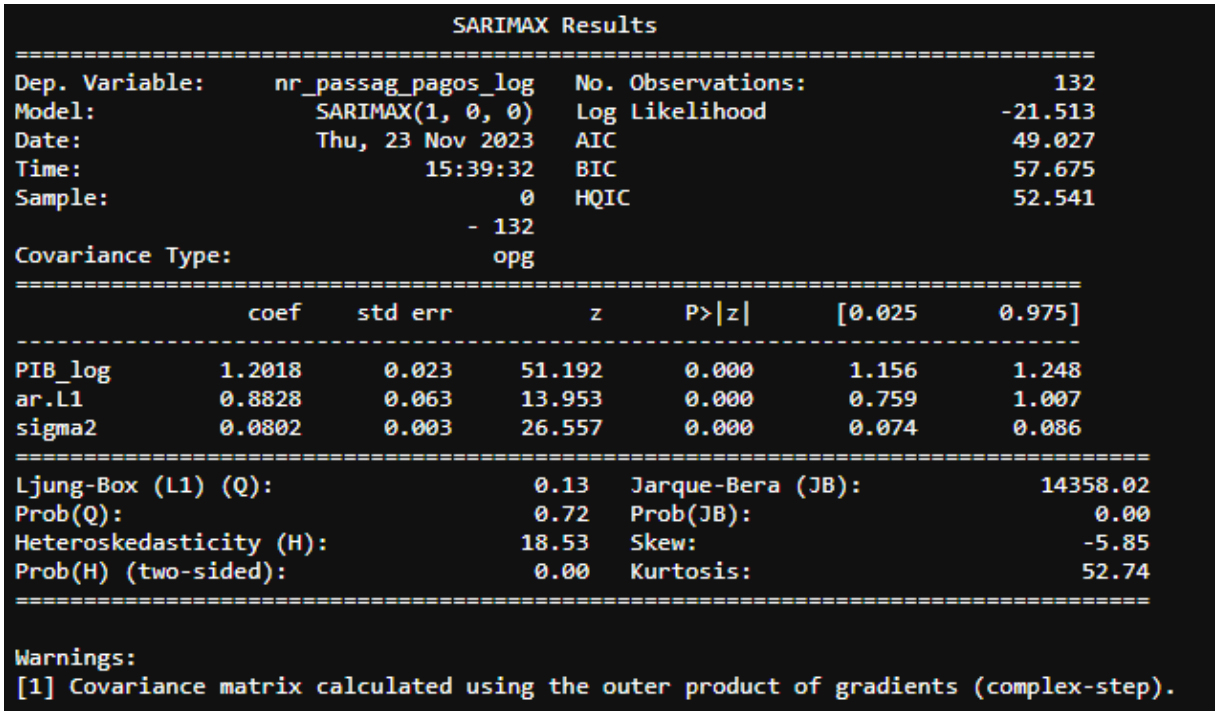
4.2.3.2 Análise do Modelo

As figuras a seguir apresentam o sumário alguns resultados do terceiro modelo:

Análise dos Testes Estatísticos:

Os resultados estatísticos do modelo são os seguintes:

- O coeficiente para $\ln(\text{PIB}_t)$ é 1.2018 com um erro padrão de 0.023, z-value de 51.192, e p-valor < 0.001, indicando uma forte influência do PIB sobre o número de passageiros pagantes.



SARIMAX Results

Dep. Variable:	nr_passag_pagos_log	No. Observations:	132
Model:	SARIMAX(1, 0, 0)	Log Likelihood	-21.513
Date:	Thu, 23 Nov 2023	AIC	49.027
Time:	15:39:32	BIC	57.675
Sample:	0	HQIC	52.541
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
PIB_log	1.2018	0.023	51.192	0.000	1.156	1.248
ar.L1	0.8828	0.063	13.953	0.000	0.759	1.007
sigma2	0.0802	0.003	26.557	0.000	0.074	0.086

Ljung-Box (L1) (Q):	0.13	Jarque-Bera (JB):	14358.02
Prob(Q):	0.72	Prob(JB):	0.00
Heteroskedasticity (H):	18.53	Skew:	-5.85
Prob(H) (two-sided):	0.00	Kurtosis:	52.74

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURA 4.10 – Sumário do terceiro modelo com apenas o PIB como variável exógena. Fonte: Autor

- O termo autoregressivo α tem um coeficiente de 0.8828 com um erro padrão de 0.063, z-value de 13.953, e p-valor < 0.001, mostrando significância estatística.
- O teste de Ljung-Box apresenta um p-valor de 0.72, sugerindo que não existe autocorrelação significativa nos resíduos do modelo.
- O teste de Jarque-Bera tem um p-valor < 0.001, o que indica que os resíduos do modelo não seguem uma distribuição normal, evidenciado também pela alta curtose de 52.74 e assimetria de -5.85.

Considerações Adicionais:

- A análise dos p-valores dos coeficientes indica que o modelo tem termos significativos que contribuem para a previsão do número de passageiros pagantes.
- A não significância do teste de Ljung-Box sugere que o modelo não deixa autocorrelação nos resíduos, o que é desejável em um modelo de séries temporais.
- A não normalidade dos resíduos, conforme indicado pelo teste de Jarque-Bera, pode implicar que transformações adicionais dos dados ou a utilização de modelos com distribuições de erro não-normais possam ser necessárias para melhorar o modelo.

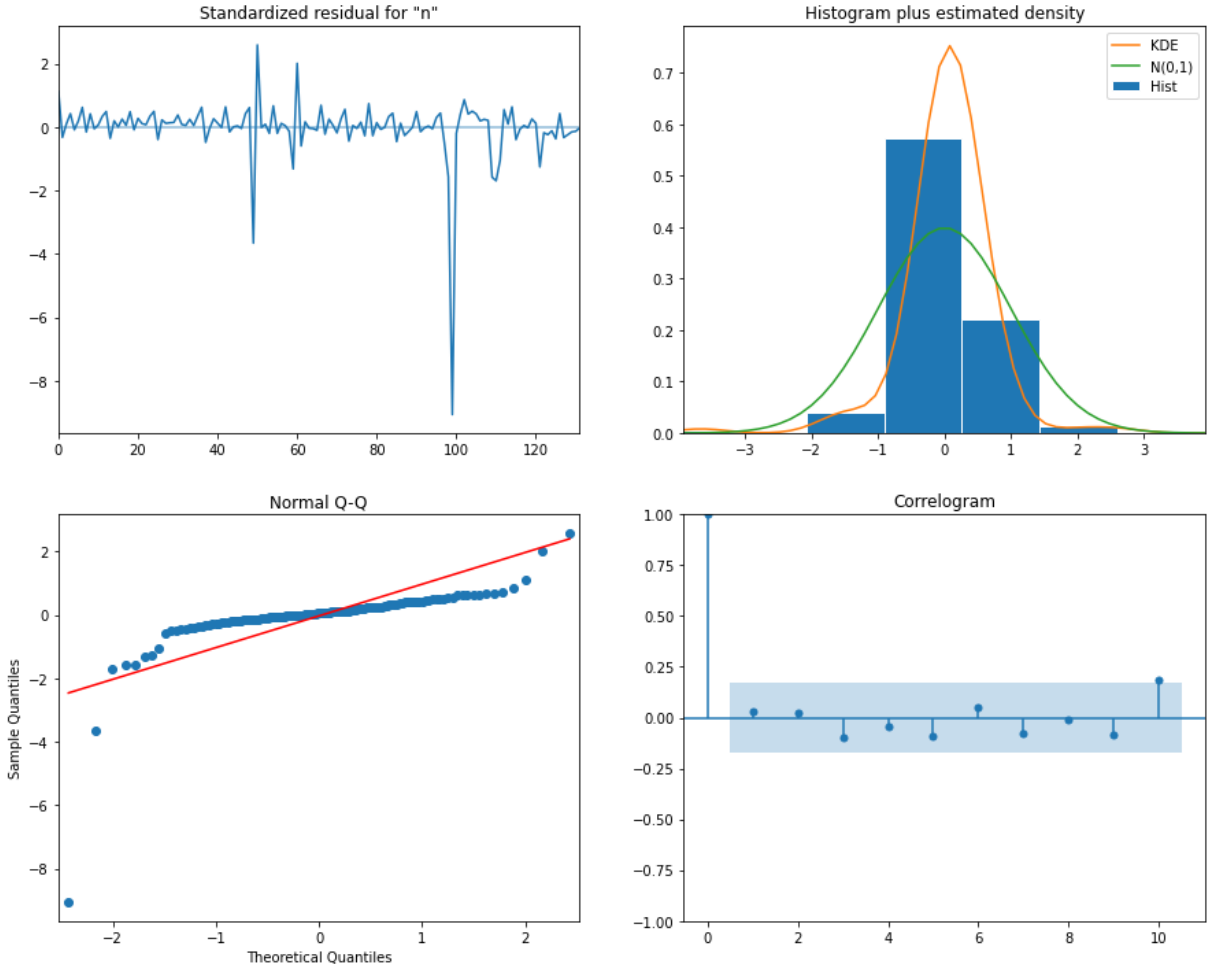


FIGURA 4.11 – Diagnóstico do terceiro modelo com apenas o PIB como variável exógena. Fonte: Autor

```
MSE: 0.044384502865504556
RMSE: 0.21067629877493232
MAE: 0.20466008653339002
```

FIGURA 4.12 – Erros obtidos no terceiro modelo com apenas o PIB como variável exógena. Fonte: Autor

4.3 Acrescentando, ao modelo com PIB, a variável obtida pelo modelo BERT como um fator exógeno

Após a realização de uma análise estatística rigorosa e a comparação dos indicadores de erro dos modelos testados, o modelo SARIMAX que apresentou a maior adequação estatística e os menores índices de erro foi o que possui ordens autoregressiva e sazonal especificadas como (p,d,q) = (1,0,0) e (P,D,Q,s) = (1,0,0,12), respectivamente. Portanto, para fins de análise subsequente, será empregado este modelo, designado como o segundo modelo no tópico de discussão anterior.

$$\ln(\text{nr\_passag\_pagos}_t) = \alpha \ln(\text{nr\_passag\_pagos}_{t-1}) + \beta \ln(\text{PIB}_t) + \Phi \ln(\text{nr\_passag\_pagos}_{t-12}) + \epsilon_t \tag{4.4}$$

4.3.1 Modelo com a primeira variável obtida pelo BERT normalizada

4.3.1.1 Representação matemática do modelo

Considere o modelo SARIMAX aplicado à série temporal do logaritmo natural do número de passageiros pagantes, com variáveis exógenas incluindo o logaritmo natural do PIB e uma variável normalizada. O modelo é especificado com uma ordem autoregressiva de um lag não sazonal e um lag sazonal, com uma periodicidade de 12. A equação do modelo é dada por:

$$Y_t = \alpha Y_{t-1} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \Phi Y_{t-12} + \epsilon_t \tag{4.5}$$

onde:

- Y_t é o logaritmo natural do número de passageiros pagantes no tempo t.
- α é o coeficiente do termo autoregressivo de primeira ordem.
- $X_{1,t}$ é o logaritmo natural do PIB no tempo t.
- $X_{2,t}$ representa a variável obtida no modelo BERT normalizada no tempo t.
- β_1 e β_2 são os coeficientes das variáveis exógenas $X_{1,t}$ e $X_{2,t}$, respectivamente.
- Φ é o coeficiente autoregressivo sazonal de primeira ordem em 12 lags.
- ϵ_t é o termo de erro aleatório no tempo t.

4.3.1.2 Análise do modelo

Os resultados indicam que o coeficiente para `variavel_bert_1_norm` não é estatisticamente significativo (p-valor = 0.978), sugerindo que essa variável não tem um efeito significativo sobre a variável dependente. Em contraste, o coeficiente para `PIB_log` é altamente significativo (p-valor < 0.001), assim como os coeficientes para os termos autorregressivos `ar.L1` e `ar.S.L12`.

Os testes de diagnóstico mostram um valor de Ljung-Box Q relativamente alto (0.27) com um p-valor associado de 0.60, indicando que não há autocorrelação significativa nos

resíduos do modelo. No entanto, o teste de Jarque-Bera revela uma distribuição de resíduos que é significativamente diferente da normal (p-valor < 0.001), e o teste de heteroscedasticidade sugere a presença de heteroscedasticidade (p-valor < 0.001). Essas questões podem comprometer as inferências feitas a partir do modelo e devem ser investigadas mais a fundo.

SARIMAX Results

Dep. Variable:	nr_passag_pagos_log	No. Observations:	132
Model:	SARIMAX(1, 0, 0)x(1, 0, 0, 12)	Log Likelihood	-17.710
Date:	Thu, 23 Nov 2023	AIC	45.421
Time:	16:38:05	BIC	59.835
Sample:	0	HQIC	51.278
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
PIB_log	1.2016	0.026	46.795	0.000	1.151	1.252
variavel_bert_1_norm	0.0086	0.305	0.028	0.978	-0.590	0.607
ar.L1	0.8634	0.090	9.646	0.000	0.688	1.039
ar.S.L12	0.2335	0.106	2.197	0.028	0.025	0.442
sigma2	0.0753	0.003	26.925	0.000	0.070	0.081

Ljung-Box (L1) (Q):	0.27	Jarque-Bera (JB):	17818.34
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	22.25	Skew:	-6.30
Prob(H) (two-sided):	0.00	Kurtosis:	58.51

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURA 4.13 – Sumário do modelo obtido utilizando PIB e a primeira variável bert criada. Fonte: Autor

4.3.2 Modelo com a segunda variável obtida pelo BERT normalizada

4.3.2.1 Representação matemática do modelo

Considere o modelo SARIMAX aplicado à série temporal do logaritmo natural do número de passageiros pagantes, com variáveis exógenas incluindo o logaritmo natural do PIB e uma variável normalizada. O modelo é especificado com uma ordem autoregressiva de um lag não sazonal e um lag sazonal, com uma periodicidade de 12. A equação do modelo é dada por:

$$Y_t = \alpha Y_{t-1} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \Phi Y_{t-12} + \epsilon_t \tag{4.6}$$

onde:

- Y_t é o logaritmo natural do número de passageiros pagantes no tempo t.

- α é o coeficiente do termo autoregressivo de primeira ordem.
- $X_{1,t}$ é o logaritmo natural do PIB no tempo t.
- $X_{2,t}$ representa a variável obtida no modelo BERT e ponderada pelo número de notícias e normalizada no tempo t.
- β_1 e β_2 são os coeficientes das variáveis exógenas $X_{1,t}$ e $X_{2,t}$, respectivamente.
- Φ é o coeficiente autoregressivo sazonal de primeira ordem em 12 lags.
- ϵ_t é o termo de erro aleatório no tempo t.

4.3.2.2 Análise do modelo

Os testes de diagnóstico indicam a ausência de autocorrelação residual significativa (Ljung-Box Q = 0.27; p-valor = 0.60). No entanto, o teste de Jarque-Bera revela que os resíduos não são normalmente distribuídos (JB = 17760.09; p-valor $<$ 0.001), e o teste de heteroscedasticidade demonstra a presença de heteroscedasticidade nos resíduos (H = 22.87; p-valor $<$ 0.001).

A variável `variavel_bert_2_norm` não é estatisticamente significativa no modelo, o que sugere que ela não contribui para explicar a variância da variável dependente e pode ser removida do modelo em futuras análises. As outras variáveis, incluindo `PIB_log`, `ar.L1`, e `ar.S.L12`, mostram-se significativas e devem ser mantidas. A não normalidade e a heteroscedasticidade dos resíduos sugerem que o modelo pode precisar de ajustes adicionais ou de técnicas de modelagem mais avançadas para melhorar o ajuste.

4.4 Considerando PIB e ICC como fatores exógenos

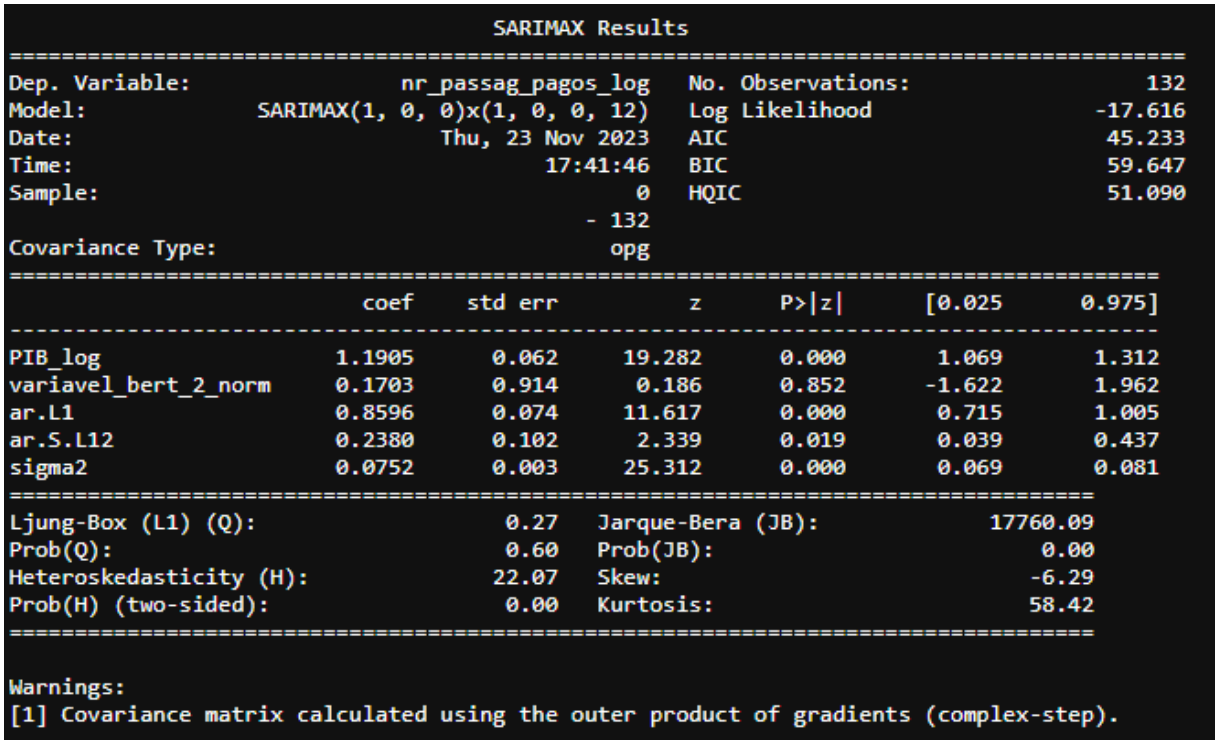
4.4.0.1 Representação matemática do modelo

O modelo SARIMAX aplicado à série temporal do logaritmo natural do número de passageiros pagantes, com variáveis exógenas incluindo o logaritmo natural do PIB e o ICC. O modelo é especificado com uma ordem autoregressiva de um lag não sazonal e um lag sazonal, com uma periodicidade de 12. A equação do modelo é dada por:

$$Y_t = \alpha Y_{t-1} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \Phi Y_{t-12} + \epsilon_t \tag{4.7}$$

ode:

- Y_t é o logaritmo natural do número de passageiros pagantes no tempo t.



SARIMAX Results

Dep. Variable:	nr_passag_pagos_log	No. Observations:	132
Model:	SARIMAX(1, 0, 0)x(1, 0, 0, 12)	Log Likelihood	-17.616
Date:	Thu, 23 Nov 2023	AIC	45.233
Time:	17:41:46	BIC	59.647
Sample:	0	HQIC	51.090
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
PIB_log	1.1905	0.062	19.282	0.000	1.069	1.312
variavel_bert_2_norm	0.1703	0.914	0.186	0.852	-1.622	1.962
ar.L1	0.8596	0.074	11.617	0.000	0.715	1.005
ar.S.L12	0.2380	0.102	2.339	0.019	0.039	0.437
sigma2	0.0752	0.003	25.312	0.000	0.069	0.081

Ljung-Box (L1) (Q):	0.27	Jarque-Bera (JB):	17760.09
Prob(Q):	0.60	Prob(JB):	0.00
Heteroskedasticity (H):	22.07	Skew:	-6.29
Prob(H) (two-sided):	0.00	Kurtosis:	58.42

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

FIGURA 4.14 – Sumário do modelo obtido utilizando PIB e a segunda variável bert criada. Fonte: Autor

- α é o coeficiente do termo autoregressivo de primeira ordem.
- $X_{1,t}$ é o logaritmo natural do PIB no tempo t.
- $X_{2,t}$ representa o ICC no tempo t.
- β_1 e β_2 são os coeficientes das variáveis exógenas $X_{1,t}$ e $X_{2,t}$, respectivamente.
- Φ é o coeficiente autoregressivo sazonal de primeira ordem em 12 lags.
- ϵ_t é o termo de erro aleatório no tempo t.

4.4.0.2 Análise do modelo

A ausência de autocorrelação residual significativa foi verificada (Ljung-Box Q = 0.00; p-valor = 0.99). Enquanto isso, o teste de Jarque-Bera revela que os resíduos não são normalmente distribuídos (JB = 12707.79; p-valor $<$ 0.001), e o teste de heteroscedasticidade demonstra a presença de heteroscedasticidade nos resíduos (H = 14.89; p-valor $<$ 0.001).

Por fim, a variável exógena acrescentada, o ICC, não se demonstrou estatiscamente significante nesse teste.

5 Considerações finais

5.1 Conclusão

Considerando a problemática de pesquisa relacionada ao aumento da demanda por passagens aéreas em todo o mundo, bem como a identificação de uma tendência de crescimento incipiente na demanda brasileira, o presente estudo empreendeu uma abordagem com o intuito de capturar uma parcela dessa variável por meio da análise de notícias que descrevem o contexto econômico brasileiro com notável sucesso.

Para atingir seus objetivos específicos, o estudo seguiu um processo em várias etapas. Primeiramente, foi construída uma base de dados robusta por meio de técnicas de web scraping, e essa base de notícias foi posteriormente submetida a um modelo pré-treinado do Google BERT, resultando na geração de variáveis relevantes. Durante a análise exploratória, identificou-se uma falha na coleta de dados em períodos anteriores a 2012, exigindo um filtro para restringir a análise aos dados coletados a partir desse ano. Essas variáveis, junto com métricas macroeconômicas pertinentes, foram então incorporadas ao modelo SARIMAX, escolhido para conduzir análises detalhadas e realizar previsões relacionadas à demanda por passagens aéreas.

No contexto da construção do modelo em questão, o primeiro passo essencial foi a incorporação do Produto Interno Bruto (PIB) como variável exógena. Esta decisão se fundamentou em estudos prévios que já haviam corroborado o significativo impacto do PIB na previsão da demanda por viagens aéreas. O PIB, enquanto uma métrica macroeconômica, é universalmente reconhecido como uma sólida representação da renda da população e de seu potencial crescimento, proporcionando insights cruciais sobre os gastos futuros com viagens aéreas.

Seguindo essa premissa, procedemos à análise de dois modelos SARIMAX, baseando-nos na exploração das funções de autocorrelação total e parcial. Inicialmente, foi adotado um modelo com os seguintes parâmetros: (p, d, q, P, D, Q, S) = (1, 0, 1, 1, 0, 0, 12). Contudo, após rigorosa análise, foi constatado que o modelo mais adequado era aquele com os seguintes resultados de parâmetros: (p, d, q, P, D, Q, S) = (1, 0, 0, 1, 0, 0, 12). Subsequentemente, esse modelo foi comparado a um modelo ARIMAX com parâmetros

semelhantes (p, d, q). A análise dos erros revelou que o modelo SARIMAX era, de fato, mais apropriado para treinamento e teste.

Em seguida, introduzimos as variáveis geradas pelo modelo BERT de maneira normalizada ao modelo. Isso incluiu tanto a média mensal da primeira variável BERT (denominada como "variável BERT 1") quanto a média mensal da segunda variável BERT, ponderada pela quantidade de notícias (conhecida como "variável BERT 2"). No entanto, os resultados de ambos os modelos não forneceram evidências estatísticas que respaldassem a utilização da análise de sentimento de notícias como um fator significativo na melhoria da previsão da demanda por passagens aéreas.

5.2 Trabalhos futuros

Para os trabalhos futuros na análise da demanda por passagens aéreas, uma série de melhorias pode ser implementada. Inicialmente, a exploração de novos modelos analíticos e preditivos é fundamental. Além dos modelos SARIMAX e ARIMAX já utilizados, a pesquisa poderia se estender para incluir técnicas avançadas como redes neurais, aprendizado de máquina e análises complexas de séries temporais. Esses modelos podem oferecer insights mais profundos e previsões mais precisas sobre as tendências do mercado de aviação.

Além disso, o refinamento no uso de tags e na filtragem de notícias é essencial para melhorar a seleção e análise dos dados coletados. Ao implementar um sistema de filtragem mais sofisticado, é possível distinguir entre notícias de maior e menor relevância, concentrando-se em tags específicas que estejam intimamente relacionadas à indústria aérea e ao contexto econômico. Isso assegura que a análise seja baseada em dados mais precisos e relevantes para o estudo.

Por outro lado, a diversificação das fontes de notícias pode trazer uma nova dimensão ao estudo. A inclusão de fontes adicionais, tanto nacionais quanto internacionais, ajudaria a capturar uma gama mais ampla de perspectivas e informações. Embora haja uma preocupação sobre a sobreposição e repetição das notícias, uma avaliação cuidadosa da unicidade das informações entre diferentes fontes pode revelar insights únicos e valiosos para a análise.

Finalmente, há um grande potencial em focar em análises regionais. Em vez de uma abordagem que considera o país como um todo, estudos detalhados sobre regiões ou estados específicos podem desvendar as dinâmicas locais que influenciam a demanda por passagens aéreas. Tais estudos permitiriam uma compreensão mais aprofundada dos fatores econômicos, sociais e turísticos que impactam diferentes áreas, oferecendo uma perspectiva mais granular e relevante para a indústria aérea. Essa abordagem regionalizada

também poderia destacar variações significativas nas tendências de viagens aéreas, que podem ser mascaradas em uma análise mais generalizada.

Referências

BAELDUNG. **Dimensionality of Word Embeddings**. 2023. Available at: https://www.baeldung.com/cs/dimensionality-word-embeddings.

BOX, G. E. P.; JENKINS, G. M. **Time Series Analysis: Forecasting and Control**. 5th. ed. [*S.l.*]: Wiley, 2015.

EISENSTEIN, J. **Introduction to Natural Language Processing**. [*S.l.*]: MIT Press, 2019.

ISHUTKINA, M. A. **Analysis of the Interaction Between Air Transportation and Economic Activity: A Worldwide Perspective**. Thesis (Doutorado) — Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.

SHUMWAY, R. H.; STOFFER, D. S. **Time Series Analysis and Its Applications: With R Examples**. 4th. ed. [*S.l.*]: Springer, 2017.

TOLCHA, T. D. The impact of economic development on domestic air traffic demand. **Elsevier**, Editora, v. 86, n. 102771, p. 20, 2020.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	27 de novembro de 2023	DCTA/ITA/TC-144/2023	53

5. TÍTULO E SUBTÍTULO:

Modelagem de demanda por passagem aérea por meio do processamento de linguagem natural

6. AUTOR(ES):

Fabio Freitas de Souza Filho

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

SARIMAX, PIB, PNL, Web Scrapping, Demanda por passagem aérea.

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Transporte aéreo; Demanda (Economia); Processamento da linguagem natural; Planejamento estratégico; Transportes.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientador: Prof. Dr. Marcelo Xavier Guterres. Publicado em 2023.

11. RESUMO:

Este trabalho de conclusão de curso investiga a integração de variáveis geradas pelo modelo de Processamento de Linguagem Natural (PLN) Google BERT em um modelo econométrico focado na previsão de demanda por passagens aéreas. Além disso, o modelo incluiu indicadores macroeconômicos, com ênfase no Produto Interno Bruto (PIB). A pesquisa iniciou-se com a coleta de notícias relevantes ao setor aéreo por meio de técnicas de web scraping, com o objetivo de criar um banco de dados para análise pelo modelo BERT.

O propósito principal era examinar se as informações extraídas das notícias, quando convertidas em variáveis pelo BERT, poderiam enriquecer as previsões do modelo econométrico. No entanto, os resultados obtidos indicaram que, apesar da metodologia inovadora e da integração de dados não estruturados, as variáveis derivadas do BERT não apresentaram significância estatística para o modelo. Isto sugere que, no contexto específico deste estudo, as nuances linguísticas e sentimentais das notícias não tiveram impacto mensurável na demanda por passagens aéreas, quando comparadas com variáveis tradicionais como o PIB.

Este achado proporciona insights valiosos para a área de modelagem econométrica, destacando a importância de avaliar a relevância e o impacto de diferentes tipos de dados. A pesquisa realça o desafio de integrar dados de PLN em modelos econométricos e sugere a necessidade de mais estudos para explorar as condições sob as quais esses dados podem ser significativos. Este estudo contribui para o corpo de conhecimento em economia e PLN, fornecendo uma base para futuras investigações sobre a aplicabilidade de técnicas de PLN em análises econômicas.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**